THE AMERICAN ECONOMIC REVIEW

March 1979

VOLUME 69, NUMBER 1

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H. Borts, Managing Editor, Box Q, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript. $15 for members, $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor

• No responsibility for the views expressed by authors in this *REVIEW* is assumed by the editors or the publishers, The American Economic Association.

Articles

Shorter Papers

THE AMERICAN ECONOMIC REVIEW

VOL. 69 NO. 2 MAY

PAPERS AND PROCEEDINGS

OF THE

Ninety-First Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Chicago, Illinois

August 29–31, 1978

Program Arranged by ROBERT M. SOLOW

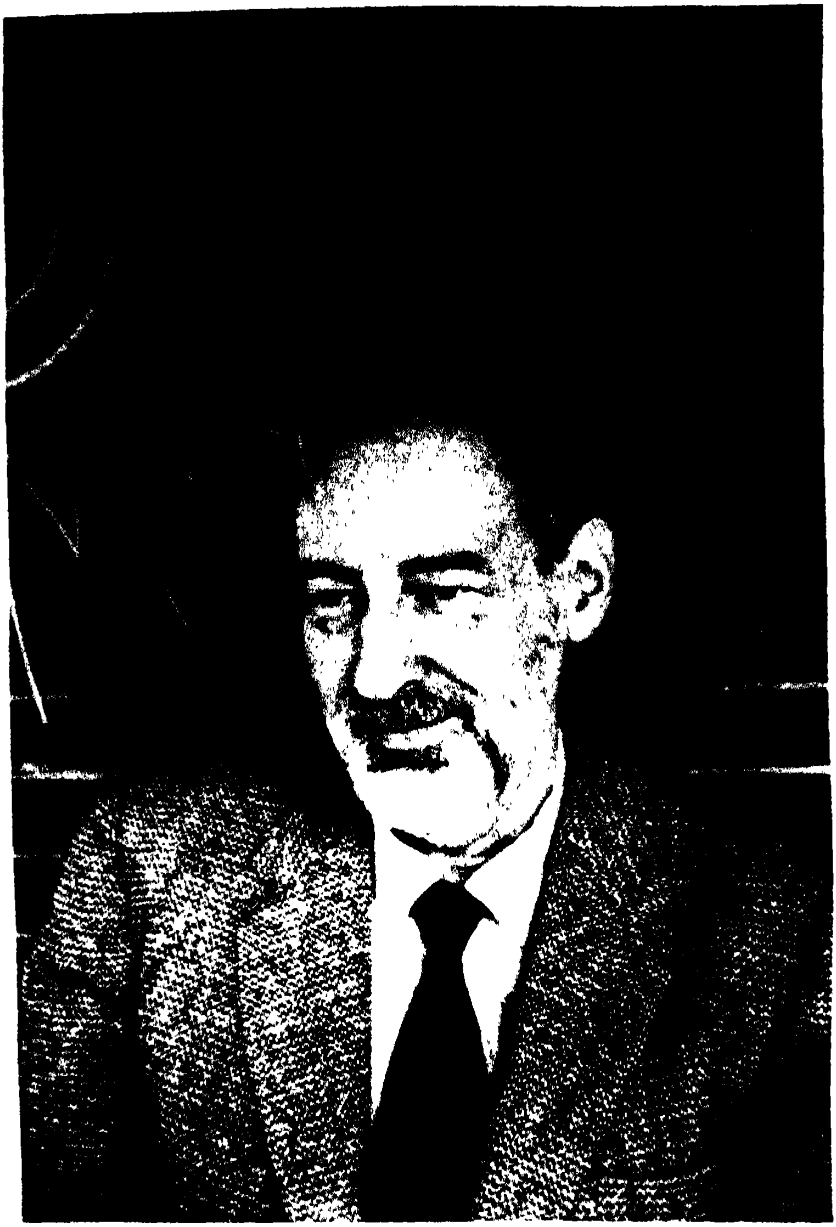
Papers and Proceedings Edited by GEORGE H. BORTS

AND JAMES A. HANSON

CONTENTS

PAPERS

Number 80 of a series of photographs of past presidents of the Association



Tjalling C. Koopmans

Economics Among the Sciences

By TJALLING C. KOOPMANS*

The title of my address implicitly assumes that economics is itself one of the sciences. I believe that to be so, and intend as I go on to indicate more fully in what sense I hold that view. However, my principal aim in choosing my topic is not that of claiming any particular status for economic analysis. Rather, I want to share with you some observations I have made over the last six years as a result of involvement in various interdisciplinary studies, through reading the reports of other such studies, or discussing them with colleagues in various fields of science.

With increasing frequency natural and social scientists are indeed finding themselves thrown together in the study of new problems that are of great practical importance for society, and essentially interdisciplinary in character. Prominent among these are problems of environmental policy, such as the protection of air and water quality. Another class of problems concerns a desirable long-range mix of technologies of energy supply, conversion and use. These two classes of problems overlap, for instance, with respect to the disposal of nuclear wastes, heat rejection to the environment, and—in the case of fossil fuels—the as yet poorly understood global and regional effects of sustained large releases of carbon dioxide into the atmosphere.

Assembled in pursuit of such studies, our interdisciplinary group soon finds that its diverse participants ask different questions; use different concepts; use different terms for the same concept and the same term with different meanings; explicitly or implicitly make different assumptions; and perceive different opportunities for empirical verification—which may lead them to apply different methods to that end. The result can be politely concealed bewilderment, possibly a suppressed surge of "we-and-they" feeling, in the worst case a growing mistrust that only time and sustained interaction can overcome.

I shall try to illustrate the difficulties of such interaction by a few examples from recent studies involving, besides economics, mostly the natural sciences and engineering. Limitations of experience, background, and time have compelled me to omit examples involving a strong participation from the other social sciences. Had my guide, mentor, and dear friend Jacob Marschak lived to give this address, and had he chosen a similar topic, the social sciences would have received an emphasis reflecting their importance to the problems of contemporary society. The writings Marschak left us, and the program and the *Proceedings* of last year's meeting of the American Economic Association, stand together as a monument to his awareness and vision of the actual and potential contributions of the social and behavioral sciences.

To prepare for the task I have set myself, I have requested and obtained interviews with a somewhat casually selected sample of natural scientists and engineers, and with a few colleagues in economics. Their responses have been drawn on in the preparation of this address, without attribution by name. I here express my, and indeed our, indebtedness for the help we have been given. Later on, I will cite some statements verbatim.

Table 1 can serve as a two-dimensional table of contents for my discussion. Three topics of study are listed on the left. On each of the three topics a recent study has

*Presidential address delivered at the ninety-first meeting of the American Economic Association, August 30, 1978, Chicago, Illinois. I am indebted to Asger Aaboe, William C. Brainard, Kenneth C. Hoffman, Alan S. Manne, William D. Nordhaus, Paul C. Nordine, Guy H. Orcutt, James Tobin, and Charles A. Walker for information, ideas, and suggestions used in writing this address. All errors are mine.

TABLE 1—ISSUES AND METHODS IN ESTIMATING BENEFITS AND COSTS

	Measures of Value			Estimation From			
Illustrative Decision Problems	*GNP* (a)	Health and Life (b)	Energy (c)	Production Process and other Technical Data (d)	Market Behavior (e)	Discounting (f)	Uncertainty (g)
(1) Helium Conservation	✓		✓	✓		✓	✓
(2) Technology Mix of Future Energy Supply and Use	✓	✓		✓	✓	✓	✓
(3) Automobile Emission Control					✓		

Note: Check marks designate issues or methods discussed or mentioned for each illustrative problem.

been made by or for the National Research Council. I will draw mostly on the first two studies and briefly mention the third.

My principal intent is not that of criticizing these studies or of evaluating their findings. I want merely to identify some of the issues that arise in their formulation, contrast responses to these issues in different professions, and comment on the methods that have been or might have been proposed or applied by the respective collaborating professions. Those issues and methods that I shall have time to refer to are set out along the top of the table. Each check mark in a cell of the table indicates that a reference is made to that issue or method in my discussion of that topic of study.

I. The Case for Helium Conservation

This study is described in the preface to the report of the Helium Study Committee as "a task that had to be undertaken quickly and completed with great speed." Likewise, on the concluding page (40), it is called a "preliminary analysis."

The principal current source of helium is as an optional by-product of the production of natural gas, in which it may occur in concentrations ranging (by volume) from 10 percent on down with increasing costs of separation. Present demand for various industrial and space uses falls below present supply, and a program of storage in the partially depleted Cliffside gas field near Amarillo, Texas, is in operation. The study is motivated by the anticipation of a substantially higher future demand.

The report of the Helium Study Committee lists, on pages 35 36, five steps that can be taken for the purpose of increasing the rate of storage. I paraphrase:

> Step i: Stop the current venting of helium which has been separated from natural gas allocated as a feedstock to petro-chemical industries. Store the helium instead.

> Step ii: Designate helium currently stored in Cliffside a "national strategic reserve" for possible major technical changes that may greatly expand future demand.

> Step iii: Reactivate presently idle separation plants to reduce the release of

helium resulting from productive combustion of the host gas, and store the helium instead.

Step iv: Build new helium separation plants on helium-rich gas streams. Store the helium.

Step v: Delay the use of helium-rich gas fields, undeveloped, and already producing.

Ultimate Step: Extract helium from the atmosphere.

The ultimate step is not included in the report of the Committee, but is mentioned in the transcript, page 135, of the Public Forum held as part of the study. It involves a process that by present technology costs a large multiple of the cost of extraction from natural gas containing .3 to .5 percent helium.

Steps i, iii, iv, and the "ultimate step" consist of successive technical process choices. Taken in that order, they correspond to the economist's notion of a long-run supply curve, indicating how the cost of each additional unit of supply is a rising step function of the cumulative supply up to that point—assuming a constant state of the technology of extraction. These steps need to be carried out only according as the expectation of demand growth becomes larger and firmer. Steps ii and v are steps whose timing should depend on additional factors besides the separation cost sequence already mentioned.

The expectation of a much larger demand well into the twenty-first century is documented, in the report and in the Forum, by a fascinating enumeration of anticipated future technologies. Many of these are based on or may utilize superconductivity, so far attained best by cryogenic techniques for which helium is the working fluid. The superconductivity is in turn expected to be applicable to a number of uses, such as power transmission with low energy loss, energy storage, and a number of applications of strong magnetic fields. Among the latter are several "technologies that either do not now exist or are in early stages of development" (pp. 13–18), such as magnetic containment for nuclear fusion reactors. Another possible application is magnetohydrodynamic (*MHD*) power generation that converts some of the energy contained in a high temperature gas stream from either a coal-fired burner or a nuclear fission reactor directly into electricity, instead of routing all the energy through a conventional steam cycle. The *MHD* development is further along in the Soviet Union. There also is a development— furthest along in West Germany and Japan—of magnetically levitated low-noise high-speed trains.

It should be added that research is in progress on the use of aluminum and possibly other materials reaching low resistivity at temperatures of 20–30°K, a range reachable using liquid hydrogen as a coolant. (See L. A. Hall, National Bureau of Standards Report, and E. B. Forsyth et al.)

From the economic point of view, the case for the helium storage program is not convincingly made either in the report or in the Forum. I have not found either cost or benefit estimates for the program. Actually, because of the importance of energy supply processes among the increased uses of helium listed above, the benefits cannot be estimated without comparable cost and fuel availability estimates for alternative energy supply and use technologies which have low or zero helium requirements. In other words, to assess the helium storage program, one also needs a long-run model of the energy sector of the economy that addresses the second decision problem of Table 1. One will, of course, also need to consider other important helium uses that are not directly energy related.

Before turning to problem (2), I draw attention to a few passages in the Helium Report that will provide background for Sections III and IV below, where I shall discuss the choice of measures of value in which to express benefits and costs. The report (p. 23) contains an important piece of information bearing on cost comparisons, to the effect that the energy requirements for

extracting helium from the atmosphere are about 1,000 times those for extracting it from natural gas containing .3 to .5 percent of helium. Large as that figure is, I shall describe later the economist's case (Section III) for including in the calculation inputs other than energy, such as that for plant and equipment, and (Section IV) for taking into account that the costs of all kinds in any required future extraction from the atmosphere will not be incurred until much later.

In fact, one statement in the report reads as a rejection of the idea that the time at which capital cost is incurred is at all pertinent. In a description of the possible role of the government in implementing the five steps, the report says: "The burden of the discount rate as a criterion of performance could be eliminated and the present debt to the U.S. Treasury written off" (p. 38). I shall explain in Section IV why I think not many economists will support the proposal to eliminate the criterion of the discount rate. Meanwhile, the statement leads one to infer that the capital cost component is not negligible as a factor in the decision.

II. Technology Mixes of Future Energy Supply and Use

My second illustration is a study that was carried out as an input to the deliberations of the Committee on Nuclear and Alternative Energy Systems (*CONAES*, in short), and of its Synthesis Panel. It is entitled "Energy Modeling for an Uncertain Future." As explained in the preface, it is a "supporting paper" published without having gone through the customary report review procedure of the Academy. While for the other illustrations I have not named authors or committee chairpersons, I should not conceal that I was the chairman of the group, called the Modeling Resource Group (*MRG*), which collectively did the work described in its Report. The group consisted mostly of economists and operations researchers, two somewhat like-minded professions. My comments on interdisciplinary interactions about the ideas and findings of the group will therefore draw on discussions with members of other professions within and outside *CONAES*.

For that purpose it may suffice to give only the briefest description of the questions addressed, the assumptions made, and the methods used. One important question arises from the fact that several competing objectives enter into the choice of a long-run energy technology mix. The net economic effect (economic benefit minus cost) of the development of a given technology mix can be estimated in a crude way, as suggested in Table 1, column (a), by its effect on the Gross National Product (*GNP*). In addition, one will also want to register risks of adverse effects such as mining accidents, air pollution, acid rain, oil spills, possible leakage from nuclear waste disposal, or diversion or proliferation of weapons-grade nuclear materials. For brevity, all such effects will be called "environmental" effects. The place in Table 1 for these impacts is column (b), tersely dubbed "health and life."

The Risk/Impact Panel of *CONAES* decided not to try to estimate money equivalents for such adverse impacts of various magnitudes. Were such estimates possible and available, then one could also define and find a balance between desired benefits and adverse "environmental" impacts that remain after scrubbers, inspectors, Civex and the like have done their jobs. Not having such estimates, the *MRG* turned the question around: Assume that tentatively chosen upper bounds are imposed on the use of technologies that have such impacts. Estimate the loss in *GNP* associated with these bounds. Then that number also places a price tag on the reduction in "environmental" impacts achieved by those bounds. Thus, even if an a priori valuation of the reduction in impacts is not available, then such a valuation is still implicit in any decision actually taken. It may help the decision makers to know these implicit price tags.

I will list only the principal assumptions made for this purpose. Numerical values were assigned to three sets of variables. As

principal exogenous, also called "realization," variables we chose:

R1. The growth rate of real *GNP*, out to 2010, in the absence of new environmental bounds on energy technologies.
R2. Capital cost levels of present and potential future energy technologies.
R3. Availabilities of oil, gas, uranium, at various costs of extraction.
R4. Long-run price and income elasticities of demand for end-use energy forms.

The "policy" variables represent the hypothetical bounds already described,

P1. Moratoria on new nuclear construction.
P2. Limits placed on output of coal and shale oil.

Forming a third category, the "blend" variables have traits of both realization and policy variables,

B1. Discount rates applied to future benefits and costs.
B2. Oil import price or quantity ceilings.

The method applied was to compare the already specified projection of a rising future *GNP* in which no bounds have been imposed on the use of energy technologies (the *base case*), with other projections in which such a bound or bounds were imposed. This procedure was carried out for each of three long-run models of the *U.S.* energy sector. The numerical inputs into the three models were the same for almost all realization and blend variables, except for the price and income elasticities of demand, which were specific to each model. Two ideas central to current economic analysis entered into this procedure. One is the use of an optimization algorithm to simulate the behavior of a competitive market economy, in any one year, and through time. The other is the use of long-run elasticities of demand. For demand by end-use consumers, these are to be based on econometric analysis of time-series and/or cross sections of income, prices, and quantities consumed. For industrial demand, a process analysis of alternative industrial energy-using processes may add valuable information.

The principal finding was the proportionally small effect on *GNP* of sizable cuts in energy use below its base case growth path. In interpreting this finding, note that the optimization procedure implies an assumption whereby the economy responds to anticipated changes by minimizing the cost of adaptation. The principal means of adaptation are changes in the type and composition of the capital equipment for the extraction, conversion, transport, and use of energy—at the regular time for replacement or earlier.

Table 2 shows the numerical results. For two models, with price elasticities of $-.25$ and $-.4$, respectively, policies entailing percentage cuts in energy use out of the base case that gradually increase to between 10 and 20 percent in the year 2010, were found to cut not more than 2 percent out of the discounted sum of annual real *GNP*, 1975-2010, and a comparably small percentage out of *GNP* for the single year 2010. The instruments of the curtailment of growth in energy use were, in row (1) of the table, the placement of bounds on specific energy supply technologies described above. In row (2), a zero-energy-growth path is simulated by the imposition of a hypothetical "conservation tax" on primary energy flows. The rate of this fictitious tax must increase as the *GNP* continues to grow in spite of the downward pull from the zero-energy-growth path.

Another finding, reported in row (2), was that for the effects of the larger cuts in energy use, the value of the long-run price elasticity of demand for energy becomes crucial. In a sensitivity analysis made with one model, a zero-energy-growth policy from 1975 on, leading to a 60 percent cut in energy use out of the base case in 2010, was found to induce only a 2 percent cut in cumulative discounted *GNP* if that price elasticity is $-.5$, but a 30 percent cut if it is

TABLE 2—ESTIMATED FEEDBACK FROM CURTAILED GROWTH OF ENERGY USE TO *GNP*, 1975–2010, UNITED STATES

Policies	Reduction out of the Base Case in: Energy Use in 2010	Discounted[a] Sum of *GNP* 1975–2010	Price Elasticity of Demand for Energy
(1) Bounds on Specific Technologies	Up to 20%	1 to 2%	−.25 or −.4
(2) Zero Energy Growth through Conservation Tax	60%	1 to 2%	−.5
		up to 30%	−.25

Source: *MRG* report, Tables III.22 and III.23.
[a]Discount Rate: 6 percent per annum.

−.25. I shall come back later to the estimation of the elasticity parameters found to have been very important.

III. Interdisciplinary Differences in Outlook

We have now assembled enough reference material for us to make a start with our main topic—the way in which differences in outlook between the disciplines affect the conduct and evaluation of joint studies.

The most significant difference between economics and the natural sciences lies in the opportunities for testing and verification of hypotheses. Jacob Marschak used to say that economists carry the combined burdens of meteorologists and engineers. Like the meteorologists, they are expected to predict the future course of important variables in their field of study. Just as engineers design more and more efficient machines, economists are also expected to improve the design of society where it affects good use of resources. But, like the meteorologist, the economist has traditionally been confined to drawing inferences from passive observations, records of data generated by the turbulence of the atmosphere or the fluctuations and trends of economic life. Finally—a very important difference—meteorologists and engineers have all the laws and measurements established by physics and chemistry available to them, fully documented by experimental tests and results.

Traditionally, economists have not searched for similar inputs from experimental or observational research of a psychological or sociological nature. In the 1950's and early 1960's they have engaged in some experimentation of their own on behavior under uncertainty and in bargaining and gaming situations. However, the findings of this work have not been put to use as premises for modeling an entire economy. For that purpose, over a few articulate protests, many economists have been satisfied to postulate simple rules of behavior by consumers and business firms. The terms "introspection" and "casual empiricism" have been used to describe the cognitive sources of these premises. In the version of the currently dominant "neoclassical" school of thought, these premises express optimizing responses of demand and supply to a uniform price system: satisfaction maximizing by consumers, profit maximizing by firms.

These premises have a certain intuitive plausibility about them. Undoubtedly, their widespread adoption has also been aided by the richness of the body of inferences one can draw from them. In fact, the premises form the logical foundation for the paradigm of neoclassical economics: the concept of an equilibrium of prices and quantities that in some way ties together the economic decisions taken by all seemingly independent agents. Conceptually the prices and the quantity responses may describe a stationary state over an extended period of time. More realistically, they may be dated variables and thus also link decisions that vary over successive periods, to sustain a moving intertemporal equilibrium.

Parenthetically, use of the term equilibrium does not imply an assumption that

the real economy actually is at any time "in equilibrium." Rather, the notion of equilibrium is a first approximation, a reference point or path, like the cycles of Ptolemy without the epicycles and the eccentricity.

If the market were to extend to all pertinent economic decisions over the entire period considered, the result of an intertemporal equilibrium would be an "efficient" path of the economy in the limited sense that no one can be made better off at any time without someone being made worse off at some time. Where market power interferes with competition, or where important economic decisions are made at government levels, the instinct of the neoclassical economist is to recommend that legislation, regulation, the use of suitable incentives, or direct government decision either restore or mimic the operation of the competitive market.

In the present context, an important trait of the neoclassical model is that it does not postulate one sole primary resource, be it labor, energy or any other, whose scarcity controls that of all other goods, and which thereby becomes a natural unit of value for all other goods. The model of production is such that—not by logical necessity, but as an empirical fact—any primary input to production can be substituted to some extent for any other. If such substitution does not take place within one-and-the-same production process, then it can still come about through suitable changes in the levels of several processes and in the inputs to these. In this view "the energy problem" is not one of just "saving energy," regardless of the cost in other resources. It is rather one of seeing to it that the increasing real cost of domestic energy extraction and supply, and the increased market power of *OPEC*, are—over time—reflected in the real prices of primary energy forms relative to other primary inputs, and thereby in different degrees in the prices of all other goods and services. In the projections described above, the energy prices are calculated so as to be in balance with an efficient path of the technology mix into the future, and thereby to induce the right amount of energy saving. In particular if, as projected by *MRG*, real prices of primary energy rise in this path, then energy use is projected to grow less than proportionally to *GNP*.

The contrary doctrine—that regardless of prices there is a persistent relationship between energy use and *GNP*—has frequently been expressed in the engineering literature. In line with this observation, the *MRG* finding of a possible small impact on *GNP* of incisive bounds on specific energy technologies led to lively correspondence and discussions with members of *CONAES*, and of its Supply/Delivery Panel, an engineering-oriented group. I should add that the *MRG* study was not the first modeling study to cast doubt on the doctrine referred to. By my knowledge the first was a study by Edward Hudson and Dale Jorgenson.

IV. Discounting Future Benefits and Costs

We are now ready for a closer look at the discounting of future benefits and costs. This practice reflects a simple technological fact combined with the paradigm of equilibrium over time. The simple fact is that—short of capital saturation—society can temporarily curtail the production of current consumption goods by transferring some factors of production to the formation of additional suitable capital goods, in such a way as to return a multiple (> 1) of the same unit bundle of consumption goods in the future. Efficient intertemporal equilibrium then demands that the present value of the goods returned to consumption be equal to that of the goods not now consumed. The quantity of the future bundle being larger, its *per-unit* present value must be correspondingly lower. In a projection that gives to one unit of the future bundle a future real market price numerically equal to its current price, a discount factor $d < 1$ must be applied to the future market price to obtain the present value, per unit of the future bundle. Given competitive markets for capital, present goods and future goods, and ignoring differences in risk, different investments bear-

ing fruit in the same future year t will tend to give rise to the same discount factor

$$d_t = \left(\frac{1}{1 + r_t}\right)^t$$

where r_t is the annual discount rate applicable to the period from year zero to year t. The usual practice in cost-benefit analyses is to assume also that r_t is independent of t, $r_t = r$, say.

This reasoning simply registers the economic accounting implications of assumed intertemporal efficiency with capital nonsaturation. To many highly educated people, there is something ethically offensive about it.

A difficult practical problem on which economists still differ among themselves is how to read a good estimate for r from capital market and other data. Different tax rates on corporate and individual incomes complicate the problem. Considering this and various market imperfections, the precluded alternative use of funds drawn upon for a public project also enters into the choice of r. I will not venture into these questions here.

Coming back to helium storage, the discounting criterion would lead most economists to recommend that those steps of the storage program be implemented for which the rate of return on the total investment (not that on energy alone) exceeds or equals the discount rate appropriate to the problem. Step i, storing helium currently vented, is likely to meet the test. The problem is to estimate which of the four or six steps would.

Two final remarks, the first added as an afterthought since August 30.

The two issues we have discussed—whether to count only energy costs or all costs, and whether or not to discount future benefits and costs—are logically distinct implications of the notion of intertemporal equilibrium. However, psychologically they are related. If one counts only energy costs, everything is expressible in equivalent Btu's, and to the physicist, steeped in the law of conservation of energy, Btu's are the same everywhere and at all times. To discount future Btu's therefore seems not just strange but outright wrong. So it is. But the economist does not discount quantities of any kind. He discounts only real values, that is, quantities (including energy) multiplied by real prices that reflect the expected balance of cost and preference as of a specified future time and beyond. It is to these prices that the discount factor is applied. I am hoping that this simple distinction may help to reduce misunderstanding between the professions.

Secondly, our reasoning has proceeded blandly as if there were no uncertainty about the outcome of the development of processes expected to be substantial users of helium. If there is considerable uncertainty, economists may want to add an allowance for risk to the discount rate. They may also wish to experiment with models in which judgmental probabilities are attached to these uncertain outcomes. This device may produce insights even if the conclusions depend on admittedly uncertain premises. A study of this kind is included as chapter IV in the *MRG* report.

V. Attaching Values to Health and Life

The question of estimating the value of health and the value of life arises mostly in contexts where either public decisions, or public monitoring of private decisions, can be shaped so as to improve health and prolong life. One example is the investment of public funds to diminish physical risks to traffic by the design of roads, bridges, turnouts, and crossings. Another is regular expenditures for traffic police, building inspectors, and other law enforcers who restrain some people from killing or hurting themselves or others by recklessness or neglect.

A common trait of these decisions is that from good experience records one may be able to estimate the years of lives saved, perhaps also of health and limbs preserved, per dollar spent on efficiently run projects or activities of this kind. Such calculations make it possible to spot discrepancies between different projects in regard to "health

and life benefits" bought per dollar spent. The ideal of equilibrium then suggests redistributing expenditures, if needed, in order to maximize total benefits from the given expenditure for protection. Valuations of health or life that have a modicum of public approval could result from such redistribution. Note that these valuations, also called *shadow prices*, are in effect set by the budgetary decision makers, whether they are aware of it or not.

After such redistribution if called for, the calculation of money values of health and life registers what in good practice we consistently spend to save a life. The process recognizes that, disturbing as it is to our sensibilities, society is being compelled by the facts of technology and behavior to set up equivalences between lives of unidentified people and bundles of goods and services implicitly of the same market or shadow value thus bracketing contemporary lives together with current goods and services in the same category of exchangeables.

The examples given so far concern small to moderate risks affecting small to moderate numbers of people, less than 100 at a time, say. Moreover, the time intervals between the decision to commit funds for the reduction of risks, the actual expenditure of these funds, and the reaping of benefits therefrom are moderate, less than twenty years, say. Finally, the problems are mostly local or national, not international in scope.

The long-run choices between energy technology mixes are different in these respects. By a gradual shift from oil and gas to coal, fossil fuels can remain a principal source of energy for countries with abundant resources of coal, especially the United States, the *USSR*, and China, for a long time to come. Intensive current discussion with regard to this option concerns the possible climatic effects of the increase in the atmospheric concentration of carbon dioxide caused by continued large-scale combustion of fossil fuels or their derivatives, alongside with world-wide deforestation. Among the large-scale effects held possible are an increase in average global temperature, entailing dislocation of agriculture depending on how each region is affected, and an increase in the level of the oceans due to the melting of polar ice not previously floating. The present state of knowledge is not such as to be ready for an assessment of these risks. New hypotheses and observations appear regularly in the pages of *Science* and other journals. So I would describe this problem as involving an unknown risk to a large number of people.

If current estimates of the capital cost of central station solar power are realistic, the principal alternative to fossil fuels for bulk power generation is nuclear power. I am not qualified to even comment on the reactor safety and waste disposal problems. I assume, however, that the developers of these technologies would classify these problems in terms of very small risks to substantial numbers of people. Perhaps this leaves as the principal concern the difficulty of keeping industrial and weapons use of nuclear materials apart. Since on this one we are all groping in the dark, I feel I should describe this aspect of nuclear technology as an unknown risk to a very large number of people.

I cannot see my way through to a calculus of the value of human lives in large numbers, that would help clarify issues of the scope of those just discussed—although estimates of numbers of lives at risk are and will remain important. These are basically problems for judgment, even though the need for making these judgments will weigh hard on the people called upon to make them. But supposing I should be wrong, let me point to one apparent paradox to be faced in any attempt to bring a calculus of the shadow price of human life to bear on problems with a long time span.

Suppose one accepts as an ethical principle that, in balancing risks to human life in the present and in the future, equal numbers of lives should receive equal weight. This would make the present value of the future human life independent of the time at which it is lived. However, we have seen that as long as capital saturation is not at-

tained, the present value of a standard bundle of goods in the future decreases as that future time recedes. Hence the present value of future life relative to that of future goods will be much higher than the value of present life in relation to present goods. It should not be inferred from this that future decision makers are assumed or advised to devote greater resources to safety and health than the present decision makers, although the future ones may well want to do this for reasons of their own. The inference is rather, I submit, that the present values I have described reflect a curious mixture of three ingredients: one intertemporal ethical rule, present preferences between consumption and protection, and an assumption about savings behavior of all generations within the next fifty years, say. Under these assumptions, sets of "present values" formed at successive points in time need not, and generally will not, be consistent with each other.

VI. The Empirical Basis of Quantitative Economics

I now go on to a discussion of the empirical basis for some of the quantitative statements that economics contributes to interdisciplinary studies. I will again illustrate this question with reference to the few studies I have chosen as examples. At the same time I will emphasize the role that the premises underlying the concept of equilibrium play in this process.

The premise of profit maximization implies a subpremise of cost minimization. I regard that subpremise as fitting reality more closely than the entire premise. It underlies the supply side of the *MRG* study of future energy technology mixes I have described.

The premise of maximization of satisfaction by the consumer can be made more plausible and more applicable by a further specification. Applied to energy, it says that successive equal additions to a consumer's annual energy end-use budget are worth less and less to him. Operationally, how much each successive addition is worth to him can be measured, for instance, by that increase in the expenditure for the rest of his consumption that he would have regarded as equivalent to each next addition to his energy consumption.

This specification implies the existence of a household demand function for energy, in which per capita demand for energy decreases as its price increases, and increases as per capita real income increases. The *MRG* extended this concept to the sum of direct and indirect demand for energy, the latter being the energy used as input to the production of all nonenergy goods, including capital goods as well. Another extension distinguishes demand for individual fuels, where the demand for one fuel increases if the price of another competing fuel goes up.

These functions are then estimated from empirical data. In the procedure followed in the model with the estimated long-run price elasticity of $-.4$ mentioned above, a parametric form of these functions was fitted to cross-section and time-series data for seven *OECD* countries, including the United States, for the period 1955–72. In the model with price elasticity $-.25$ the estimation procedure was not stated with comparable explicitness. In both models the estimated long-run demand functions, written with price as a function of quantity, were then integrated to estimate the benefit from the consumption of energy in all forms.

By comparison, the empirical basis for the production side is more direct. Each of the various competing energy producing, converting, and using processes is represented by constant ratios of inputs to outputs, reflecting operating experience where available, or based on estimates of such ratios and of future availability dates for processes not yet developed. For instance, process estimates for the years 1985 and 2000 were drawn upon in estimating the elasticity of substitution between electric and nonelectric energy in the second of the two models just discussed. This did constrain but not by itself imply numerical estimates of the elasticities of demand for

energy, whether in toto (given as −.25), or for the two components.

This completes my description of the empirical basis for the *MRG* procedures and the premises on which they rest. I want, in passing, to draw attention at this point to the econometric aspects of another study, designed to estimate perceived benefits of air quality improvements from residential property values in areas with different air quality. The study is entitled "The Costs and Benefits of Automobile Emission Controls." It draws on a body of econometric work in which property values are related to various characteristics of the site and the neighborhood, including air quality and other environmental amenities, and the income of the household.

The foregoing examples lead me to some broader remarks on the empirical basis of quantitative economic knowledge in general, not limited to the type of studies we are here mostly concerned with.

In all formal procedures involving statistical testing or estimation, there are explicitly stated but untested hypotheses, often called "maintained" hypotheses by statisticians. In the econometric studies we have here considered, the "premises" already discussed play that role. More in general, any statement resulting from such studies retains the form of an "if ... then ..." statement. The set of "ifs," sometimes called "the model," is crucial to the meaning of the "thens," usually but somewhat inaccurately called the "findings." For instance, in fitting demand relations, the principal maintained hypotheses specify the variables entering into these relations, and possibly other variables with which these variables are in turn linked in other pertinent relations.

The "if ... then ..." statements are similar to those in the formal sciences. They read like logical or mathematical reasoning in the case of economic theory, and like applications of statistical methods in the case of econometric estimation or testing. The heart of substantive economics is what can be learned about the validity of the "ifs" themselves, including the "premises" discussed above. "Thens" contradicted by observation call, as time goes on, for modification of the list of "ifs" used. Absence of such contradiction gradually conveys survivor status to the "ifs" in question. So I do think a certain record of noncontradiction gradually becomes one of tentative confirmation. But the process of confirmation is slow and diffuse.

For some purposes, and at considerable expense, short cuts can be made to diminish the dependence on untested "ifs." I am speaking of systematic experiments such as the so-called negative income tax experiment conducted in New Jersey over the period 1968–72, and followed by similar income maintenance experiments in other areas of the United States. If one wants to know whether income maintenance payments to families near the poverty line have a disincentive effect, or no effect, or even a positive incentive effect on labor supply, one does not need to have a pretested theory as to what, if anything, the family is maximizing. Instead, one can make such payments to a sample of families and compare its behavior with that of an unaided control group. This is what the New Jersey experiment did. In one category of families where the numbers spoke rather clearly—white husband-and-wife whole families—the effect on labor supply was found to be negative, moderate but statistically significant, and with the effect on the husband's labor supply smaller than that on the wife's. In addition, much was learned about the design, conduct, and evaluation of such experiments for use in later studies.

There have not been many such experiments on a scale needed to obtain statistically significant outcomes. Moreover, they have been limited to questions of great and urgent policy importance. Meanwhile, we do need to find ways in which verification of the premises of economics, through cumulative econometric analyses and through experiments that find a sponsor, can be pursued.

I have not found in the literature a persuasive account of how such confirmation of premises can be perceived and docu-

mented. How do we keep track of the contradictions and confirmations? How do we keep the score of surviving hypotheses? And what are we doing in those directions? The same questions have been raised before, among others by my predecessors, Wassily Leontief and Robert Aaron Gordon, and good and bad examples of concern and unconcern were referred to by both of them. Meanwhile, unresolved issues, sometimes important ones from the policy point of view, and mostly quantitative ones, drag on and remain unresolved. Do they have to?

With one exception I am aware of, even our best college-level introductory texts of economics do not press these questions. They teach good reasoning, and describe the views of leading minds and schools of thought, present and past, in the field. Texts in econometrics teach with great care how to test assumptions and to estimate parameters, duly emphasizing the crucial role of the models. What is also needed is to teach the tested and confirmed statements.

VII. Aphorisms on Interactions

After all I have said about the need for empirical validation, I owe you a brief report on my own casual-empirical sample study of the difficulties of interaction between scientists, engineers, and economists, as seen by participants in joint studies. Rather than classifying and tabulating the views expressed, I shall let the respondents speak for themselves. The following is a selection (by me) of statements, drawn from my notes, that carried the most punch.

A physical scientist: "Economists are technological radicals. They assume everything can be done."

A geologist: "Economists have been too enthusiastic about deep sea mining. They think there is more than there is, that it is easier to get up than it is, and easier to process than it is."

A development economist: "Scientists think big. Economists are marginalists. Scientists don't think in terms of opportunity cost."

An engineer: "Economics is not dismal but incomplete. The things missed are very important."[1]

A life scientist: "Market imperfection is more widespread than economists care to admit."

An economist: "Where economists see the invisible hand guiding the market place to produce pretty good outcomes, scientists see only chaos."

An engineer: "The economic motive is overrated."

A psychologist: "All the conclusions that are drawn from the assumption of rationality can also be drawn from assumptions of adaptive behavior."

A life scientist: "Economists have great skill in handling data. However, they tend to ask only for data, not for concepts and ideas. Drawing up a model is an interdisciplinary task."

An engineer: "Economists often use smooth production functions even when engineers might be reluctant to do so."

A life scientist: "Many scientists do not understand discounting."

An engineer: "Economics is the Thermodynamics of the Social Sciences. Everything is deduced from a few simple postulates without the necessity for knowing detailed mechanisms."

VIII. Final Remarks

After this instructive intermezzo, allow me a few final words. I will not be able to match the brevity and incisiveness we just savored. However, I do look on the collaboration of the diverse professions involved in the newly discovered joint problems as an important development. To economists it is a new challenge and a new frontier. Among the problems themselves are some of great importance, nationally and internationally. They deserve the best effort and talent that can be brought to bear, within and across the disciplines.

An important talent requiring cultivation is skill in communication between dis-

[1]The reference is to the need to fit environmental protection into economic analysis

ciplines. We should begin with the defusing of jargon. Perhaps some terms should be explained at first use. To the physicist who has used calculus on problems going back to Isaac Newton, it is unexpected to learn that everything called "marginal" is a first derivative of something. It appears natural to him, however, to learn that an "elasticity" is the dimensionless slope of a curve plotted on double-log paper. There is more trouble lying in wait with "externalities," an institutional concept presupposing private property, or at least an accountability for private or public production or household decisions that is dispersed over individuals and organizations. If we will be more forthcoming with explanations of our cherished terms, our science colleagues may be more inclined to help us out with "entropy," which to me is a more difficult concept than anything economics has to offer.

A more serious problem is that, while our universities are the principal training ground for future scientists of all kinds, they do not seem to be the best place for gaining experience in interdisciplinary interaction. I believe that the root of the difficulty lies in the procedures for academic appointment and promotion. The initiative, the decisive first step, is usually taken in the department of one's own discipline. Young faculty members must prove their worth first to their senior colleagues in the field they are identified with. A joint appointment holds somewhat less promise as a stepping stone to tenure. Even our graduate students are already aware of these factors.

The increasing demand for the contributions of interdisciplinarians may gradually break the barriers down. Progress will be slow unless university faculties and administrations perceive the problem. Once they do, the irrepressible curiosity and venturesomeness of our undergraduates will provide a point at which to start and from which to build up.

REFERENCES

E. B. Forsyth et al., *Underground Power Transmission by Superconducting Cable*, Brookhaven National Laboratory 50325, Mar. 1972, esp. ch. IV.

L. A. Hall, "Survey of Electrical Resistivity Measurements on 16 Pure Metals in the Temperature Range 0 to 273°K," Nat. Bur. of Standards, Tech. Note 365, Feb. 1968, esp. p. 20.

E. A. Hudson and D. W. Jorgenson, "Economic Analysis of Alternative Energy Growth Patterns, 1975–2000," Appendix F, pp. 493–511, in *A Time to Choose*, report by the Energy Policy Project of the Ford Foundation, Cambridge, Mass. 1974.

Helium Study Committee, "Helium: A Public Policy Problem," National Resource Council, 1978.

Modeling Resource Group, "Energy Modeling for an Uncertain Future," report prepared for the Committee on Nuclear and Alternative Energy Systems, National Academy of Sciences, 1978.

Report prepared for the Committee on Public Works, U.S. Senate, "The Costs and Benefits of Automobile Emission Controls," a report by a committee for the National Academies of Sciences, and of Engineering, Sept 1974, serial no. 93-24, ch. IV.

Fertility, Women's Wage Rates, and Labor Supply

By BELTON M. FLEISHER AND GEORGE F. RHODES, JR.*

Development of the "new home economics" has emphasized the mutual determination of family size, or fertility, and the labor force behavior of married women. The following questions arise in this context: How is a mother's labor market opportunity set influenced by devotion of her time to bearing and rearing children? How do wife's market wage rate and family income affect the number and "quality" of children desired? How do the family's demand for children and the wife's labor market opportunities interact to determine the wife's lifetime labor supply? Answering these questions requires a multiple equation family model of fertility, child quality, wage rates, and labor supply. This paper presents estimates of such a model using disaggregate data of the National Longitudinal Surveys (*NLS*). Only two previously published papers present estimates based on a comparable approach, but both studies (see Glen Cain and Martin Dooley; Marc Nerlove and T. Paul Schultz) use data aggregated across geographical locations, rather than individual data. Moreover, these earlier papers differ from ours in model specification.

The estimates we obtain for our family model provide answers to the three questions posed above. These are of concern from a theoretical point of view, as we are interested in whether the family's demand for the quantity and quality of children is negatively related to their price, and the ability of the human capital-household production approach to explain interrelated forms of household behavior. Moreover, our results can be related to those of earlier studies of fertility, female labor supply, and female earnings, most of which used single equation methods. These earlier studies represent attempts to answer a number of interesting questions on which the present study provides further evidence: Does the wife's labor market experience increase her labor market earning power, and is this influence tempered by devotion of time to household production, especially rearing children? (See Jacob Mincer and Solomon Polachek.) Despite the common observation that wealthier families have fewer children, are children actually a normal commodity when the correct measure of their price is held constant? (See Gary Becker and many subsequent studies.) If so, is the effect of income on the quantity of children demanded small relative to the (opposite) effect of the price of children? What is the magnitude of the income elasticity of demand for children relative to that for child quality when the correct price measures are held constant?[1] (See Becker; Mincer, 1963; Dennis DeTray; Robert Willis; Becker and H. Gregg Lewis.) The presence of children has a pronounced negative influence on current female labor force participation in single equation models. (See Mincer, 1962; Cain, 1966; Bowen and Finegan, 1969.) Does this influence persist in a multiple equation, lifetime context? Can the remarkable histori-

*Department of economics and Center for Human Resource Research, Ohio State University, and department of economics, Colorado State University, respectively. Valuable comments were received from seminars at the National Bureau of Economic Research-West and the workshop in Applications of Economics at the University of Chicago; and from Donald Parsons, Steven Sandell, Timothy Carr, and an anonymous referee. We wish to acknowledge the invaluable assistance of E. J. Honton and Philip Mehall. This research was supported in part by grants from The National Institute of Child Health and Human Development and The Instruction and Research Computer Center of Ohio State University.

[1]For simplicity, we shall refer to these income elasticities as "true" income elasticities.

cal increase in the labor force participation of married women be understood in terms of a response to market earning opportunities and family income levels? Past research offers widely contrasting answers to this question. (See Mincer, 1962; Cain; William Bowen and T. Aldrich Finegan; Schultz.)

The specification of our model differs from earlier multiple equation models in the following ways. We specify that parental demand for child quality affects the price of adding an additional child to the family; conversely, the number of children affects the price to parents of increasing the quality of each child if all children are treated alike. Since the cost of parental time inputs varies across families, this approach leads to an interactive specification in which the product of a proxy for the cost of parental time and the amount of child quality demanded represents the price of number of children, and the product of time cost and the number of children represents the price of a unit of child quality. We focus on wife's lifetime labor supply as influenced by fertility decisions, price, and income variables, rather than current (census week) labor force participation, which data limitations forced on earlier studies. Our measure of fertility is total number of children living at home or elsewhere and is a much closer approximation to a measure of completed family size than is Nerlove and Schultz's measure of current birth rates by geographical area. Our fertility equation includes the budget constraint variables representing family income and the price of child services, but it does not include a measure of wife's labor supply as do the studies of Nerlove and Schultz and Cain and Dooley. We believe our formulation permits a clearer understanding of the total effect of budget variables on family size decisions.

Much of the advantage possessed by the present study over earlier attempts to estimate multiple equation family models derives from the especially rich body of information contained in the *NLS*. Although we are forced to use proxies in several instances, on the whole we are able to measure crucial variables more satisfactorily than was the case in previous studies. Information on work history permits us to estimate both a human capital wage function and a labor supply equation without excluding wage information on more than half of the sample who were not working when surveyed. This reduces the likelihood of censorship bias in these relationships. Variables pertaining to three generations can be linked together for each family because of the *NLS* sample design. This feature, in conjunction with the disaggregate nature of the data, assures that information for each observation all pertains to the same household or to related households. The information on parents' family background yields exogenous variables which increase our ability to identify the structural equations of the model.

In Section I, we outline the theoretical underpinnings of the model, describe the data base and variables to be used, and discuss the econometric properties of our estimation procedure. Section II summarizes the estimates and compares them with those reported in previous research on fertility and the labor force behavior of married women.

I. The Model

A. *Theoretical Framework*

Following Willis we assume that the family behaves as though it were maximizing the family utility function

$$(1) \qquad U = U(N, Q, S)$$

where N, Q, and S represent the number of children, child "quality," and nonchild sources of satisfaction, respectively. Q and S are nontraded commodities produced in the home according to linear homogeneous production functions:

$$(2) \qquad Q = f(t_c/N, x_c/N)$$

$$(3) \qquad S = g(t_s, x_s)$$

where t_i and x_i, $i = c, s$, represent time and goods, respectively, devoted to Q and S

production. Equation (2) can be multiplied by the family's total number of children N, yielding

$$(4) \qquad C = NQ = f(T_c, x_c)$$

where C represents the aggregate amount of child quality produced by the family, or "child services." It is assumed that joint production of C and S does not occur.

It is also assumed that the husband contributes no time to t_c or t_s (i.e., he works "full time" in the market). Thus, the family faces the following constraint with respect to its purchases of the market good $x = x_s + x_c$:

$$(5) \qquad Y = H + wL = px$$

where p = the price of $x = 1$, H represents the family's nonlabor lifetime income or wealth plus the husband's lifetime earnings, w is the wife's average lifetime wage, and L is the amount of time she works in the labor market after marriage. The family also faces the following constraint with respect to the wife's time:

$$(6) \qquad T = t + L$$

where $t = t_c + t_s$ and T, the wife's lifespan following marriage, is assumed to be exogenous.

Analysis of the model is facilitated by reformulating the family's budget and time constraints in terms of "full" income and the shadow prices of the commodities C and S as follows:

$$(7) \quad I = \pi_c NQ + \pi_s S = \pi_c C + \pi_s S$$

where π_c and π_s are the shadow prices of C and S, respectively. Since it assumed that the production of C is relatively intensive in mother's time, π_c and w are positively related, and w is used as a proxy for π_c in our empirical work. The demand for N, Q, and S can then be expressed as functions of I, π_c, and π_s; empirical analysis is carried out by letting π_s be the numeraire and relating π_c and I to w and H, respectively.

Empirical analysis is complicated by the interdependence of crucial variables in the model. Following the work of Hendrik Houthakker, papers by Henri Theil, Becker, Willis, and Becker and Lewis have shown that the full price of an additional child is $\pi_c Q$ and that of an additional unit of child quality is $\pi_c N$. Therefore, the price of N is a function of Q, while the price for Q depends on N. If both Q and N are normal commodities, then failure to correctly specify the price of children as $\pi_c Q$ and the price of child quality as $\pi_c N$ will lead to downward biased estimates of the income elasticity of demand for N and Q. Furthermore, w must be treated as endogenous because a woman's investment in her own capital, especially through on-the-job training and also through schooling, will be affected by her desired and actual number of children. Thus, the mother's labor market experience and N are endogenous variables, both influencing her wage. Clearly, the inclusion of N in ordinary least squares (OLS) regression estimates of labor force participation equations for married women is likely to result in biased estimates of the effects of w, H, and N on labor force participation. (This point has been emphasized by Schultz.)

B. *Empirical Specification of the Model*

In this section we discuss the empirical specification of a one-period, lifetime family model. The model is estimated by means of instrumental variables with data for black and white families from the *NLS*s of women ages 30–44 and young persons ages 14–24. We are forced to use the instrumental variables procedure because a different number of observations is available to estimate each equation, precluding the use of two- or three-stage least squares. The reason for the disparity in sample sizes is due to missing data for some variables and, more fundamentally, to the out of school *child* being the unit of observation for equation (10), the child quality equation, while the *household* is the unit of observation for equations (8), (9), and (11) below. Table 1 defines the variables used. Exogenous variables used only in the instrumental variable equations are listed in the Appendix.

TABLE 1—VARIABLE DEFINITIONS

Endogenous Variables	
WM	Wife's average lifetime wage[a]
N	Number of children ever born, 1967
N^*	Number of children ever born, 1967 for women ages 40–44 only
Q	A child's wage rate on current or last job (child quality)[a]
R	Proportion of years mother has worked at least six months since leaving school
R^*	Proportion of years mother has worked at least six months since leaving school for women ages 40–44 only
R'	R multiplied by number of years since leaving school (i.e., a measure of labor market experience in years)
Exogenous Variables	
WF	Father's wage rate at age 40[a]
A	Mother's age in 1967
SF	Highest year of school completed by father
SM	Highest year of school completed by mother
B	Dummy variable = 1 for blacks

[a]For further information, see the Appendix.

As an approximation to the hypothetical model deduced by maximizing (1) subject to (2), (3), and (7), we use the following empirical model for purposes of estimation:

$$(8) \quad \ln WM = \alpha_{10} + \alpha_{16} WF + \alpha_{19} SM + \alpha_{110} B + \alpha_{111} R' + \alpha_{112} N \cdot R'$$

$$(9) \quad N = \alpha_{20} + \alpha_{25} WM \cdot Q + \alpha_{26} WF + \alpha_{27} A + \alpha_{28} SF + \alpha_{210} B$$

$$(10) \quad Q = \alpha_{30} + \alpha_{34} WM \cdot N + \alpha_{36} WF + \alpha_{39} SM + \alpha_{310} B$$

$$(11) \quad R = \alpha_{40} + \alpha_{41} WM + \alpha_{42} N + \alpha_{43} Q + \alpha_{46} WF + \alpha_{47} A + \alpha_{49} SM + \alpha_{410} B$$

The instrumental variables coefficient estimates for equations (8)–(11) are consistent, by their construction. However, they are less efficient than two-stage or three-stage least square estimates would be if they were obtainable.

Equation (8) is a human capital wage equation, the arguments representing variables which influence the productivity of market worktime. We justify the semilog form on grounds of its general acceptance in the human capital literature. The expected signs of the coefficients are $\alpha_{16} > 0$, $\alpha_{19} > 0$, $\alpha_{110} < 0$, $\alpha_{111} > 0$, and $\alpha_{112} < 0$. Mother's market earning power is hypothesized to be enhanced by formal schooling (SM) and on-the-job training (OJT). The term $N \cdot R'$ is included because OJT is only imperfectly captured by R', years of labor market experience. What we would also like to know is the *intensity* of training during time spent working in the labor market. It is our maintained hypothesis that mothers who bear a relatively large number of children will devote a relatively small proportion of their market work time to training activity, especially during the years before they have acquired their desired number of children (which covers the period over which R' is measured for most of our sample). It is neither in the mother's nor an employer's interest to devote resources to job-specific training if time on any given job is expected to be short. Since rearing children requires that some time be spent outside the labor force associated with each birth, during which general training may depreciate, there is less incentive to acquire general training as well when family size is large (see Mincer and Polachek).[2]

Opinion is divided on whether WF should be included in an equation such as (8) (see James Heckman; Schultz). We believe WF should be included, because family income available if the wife does not work facilitates waiting and/or searching for a desirable job, resulting in higher observed values of WM.

Equations (9) and (10) are the demand functions for number and quality of children, respectively. From theoretical considerations it is clear that equations (9) and (10) should be of the general form

$$(9') \quad N = N(I, \pi_c N, \pi_c Q)$$

[2]N may also reflect time out of the labor force and resulting human capital depreciation.

(10′) $Q = Q(I, \pi_c N, \pi_c Q)$

Adopting this specification could introduce nonlinearities in the variables, complicating estimation considerably.[3] If equations (9′) and (10′) were specified to be *log*-linear, we could write, for example,

$$(9'') \quad lnN = ln\,\alpha_0 + \alpha_1 ln\pi_c N + \alpha_2 ln\pi_c Q + \alpha_3 lnI$$

which could be estimated by conventional techniques as

$$(9''') \quad lnN = ln\frac{\alpha_0}{1-\alpha_1} + \frac{\alpha_1+\alpha_2}{1-\alpha_1} ln\pi_c + \frac{\alpha_2}{1-\alpha_1} lnQ + \frac{\alpha_3}{1-\alpha_1} lnI$$

This would, however, require the undesirable step of adding one to all variables which can take on zero values: N and control variables such as SF and SM. Moreover, the income elasticity (α_3) is unity, and experimentation with various alternative approaches all yield considerably smaller estimates. Given the nature of our data and the well-known sensitivity of results to specification and statistical procedure in non-linear estimation, we believe that the lesser of evils is to assume that the coefficients of $\pi_c N$ and $\pi_c Q$ in equations (9′) and (10′), respectively, are zero and to estimate (9) and (10) as specified. Equation (9″) and its counterpart (10″) might also be approximated with arithmetic values for the variables, using elasticities calculated at the means to compute the parameters of interest. In order to test the sensitivity of our results to the deletion of $\pi_c N$ and $\pi_c Q$ from equations (9′) and (10′), we have estimated such an approximation, and the results are presented below.

The expected signs of the coefficients in (9) are $\alpha_{25} < 0$, $\alpha_{26} > 0$, $\alpha_{27} > 0$, $\alpha_{28} < 0$, and $\alpha_{210} \gtrless 0$. As shown above, WM is positively associated with π_c, the price of "child services." The variable Q is a measure of child quality, and $WM \cdot Q$ is therefore taken to represent $\pi_c \cdot Q$, the price of N. Of course the inequality $\alpha_{25} < 0$ is expected to hold only insofar as the income effect associated with changes in WM is dominated by the substitution effect of changes in $WM \cdot Q$. The *ceteris paribus* income effect on the demand for number of children (α_{26}) is assumed to be positive.

The sign of α_{27} is assumed positive because, *ceteris paribus*, greater age increases the probability that a child will have been born. There may be offsetting cohort effects, however, particularly since the years of most frequent childbirth among women 30–44 in 1967 encompass the beginning, peak, and waning of the postwar "baby boom."

The choice of SF vs. SM for inclusion in equation (9) may appear somewhat arbitrary, since SM is included in equation (10), but not SF. Father's schooling SF is assumed to reflect contraceptive knowledge, implicitly raising the cost of N relative to other commodities. It is true that SM may also be used to reflect contraceptive knowledge (see Robert Michael) and both SM and SF may reflect efficiency in child rearing as well as dimensions of desired child quality not adequately captured by the admittedly imperfect proxy for Q.

Unfortunately, inclusion of both SM and SF tended to raise standard errors without increasing $\bar{R}^2$. Because of this effect of collinearity, we had to choose between the two variables in our empirical work, and the choice for SF in equation (9) was made because it yielded a higher $\bar{R}^2$ and lower standard errors on all coefficients relative to the magnitude of their regression coefficients.

The variable B has been included to reflect any *ceteris paribus* cultural differences between whites and blacks in desired family size.

Equation (10) is the demand equation for child quality. We feel that the expected market wage of a child reflects an important dimension of parents' goals for their children, to which other possible dimen-

[3]We are indebted to Nicholas Kiefer for emphasizing these points to us.

sions of quality such as schooling, absence of a delinquency record, and so on are better viewed as inputs. Since age data are available mainly for out-of-school persons, there is possible selectivity bias in estimating equation (10), just as there is possible bias from inclusion of Q as a right-hand variable in equations (9) and (11). Aspects of this bias are treated at some length in Fleisher. Other proxy variables for child quality—particularly focusing on various dimensions of schooling—were tried, but the empirical results differed too little from those reported to warrant their inclusion here.

The expected signs of the regression coefficients are $\alpha_{34} \gtrless 0$, $\alpha_{36} > 0$, $\alpha_{39} > 0$, $\alpha_{310} \gtrless 0$. While the rationale underlying these hypothesized signs is similar to that for equation (9), we are not as confident that we have not omitted relevant productivity variables in equation (10). The variable SM is included to reflect mother's efficiency in producing child quality (see C. Russell Hill and Frank Stafford; Arleen Leibowitz; Fleisher). Nevertheless, parents who themselves have achieved a relatively high level of economic success can be expected to be relatively efficient in helping their children achieve these goals. Therefore, it is probably difficult to separately identify price, income, and efficiency effects in the quality demand function. Since WM may reflect mother's efficiency in the production of child quality as well as the price of her time, we do not feel confident in hypothesizing an unambiguous sign for α_{34}.

Equation (11) models the mother's lifetime labor force decision. The hypothesized signs of the regression coefficients are $\alpha_{41} > 0$, $\alpha_{42} < 0$, $\alpha_{43} < 0$, $\alpha_{46} < 0$, $\alpha_{47} \gtrless 0$, $\alpha_{49} \gtrless 0$, $\alpha_{410} > 0$. The rationale underlying these hypotheses is based on the view of the household as a producing unit using the wife's time as a principal input. The variables N, Q, and WF reflect the scale of household production. With N, Q, and WF held constant, α_{41} reflects the output-constant demand for mother's time in the home plus any income effect of WM on the demand for commodities other than child services. The variable age is included to allow for the tendency of mothers to reenter the labor forces as their children grow up, although once again cohort effects may also be reflected in α_{47}. The sign of α_{49} is ambiguous because SM is assumed to reflect efficiency in home production, taste for market work, and access to relatively pleasant jobs. There is evidence that *when children are present*, SM reduces R, *ceteris paribus* (see Leibowitz; Hill and Stafford; Yoram Ben-Porath; Malcolm Cohen, S. Rea, and Robert Lerman). At other stages of the life cycle, however, women with higher schooling attainment tend to participate in the labor force to a greater extent than other women.

On the basis of prior studies, black women are expected to spend more time in the labor force than whites, *ceteris paribus*. The most likely explanation is the greater reliance in black families upon the wife's earnings because of labor market difficulties faced by black men.

II. Estimates

Estimates of the equations (8)-(11) are presented in Table 2.[4] The sample excludes families of women who were not married and living with their husbands in 1967 and whose race is other than black or white. Since work with separate data for blacks and whites suggested substantial differences between the two races, we focus on the results for whites only (shown in the lower half of Table 2) in the discussion that follows.[5] The results compare reasonably well to the hypothesis presented in Section I, but there are a few surprises.

The estimate of equation (8) conforms quite closely to our hypotheses. A one dollar increase in husband's hourly wage is associated with a 5.2 percent increase in the wife's average lifetime market earning power. An additional year of schooling in-

[4]The instrumental variable equations are available from the authors on request.

[5]A paper containing comparable results for blacks is available from the authors on request.

TABLE 2—ESTIMATES OF EQUATIONS (8)–(11)

Equation	Dependent Variable	(1) *WM*	(2) *N*	(3) *Q*	(4) *WM·N*	(5) *WM·Q*	(6) *WF*	(7) *A*	(8) *SF*	(9) *SM*	(10) *B*	(11) *R′*	(12) *N·R′*	$\bar{R}^2$	*n*
(8)	*lnWM*						.018 (1 5)			.032 (6.8)	−.18 (5.0)	054 (8 5)	− 020 (9.5)	.53	2674
(9)	*N*					− 00035 (8.5) *− 75*	64 (5 1) *60*	− 0060 (0 7)	−.13 (5.5)		.36 (2 7)			10	3394
(10)	*Q*				000016 (0.2) *.06*		17 (1 9) *18*			023 (0.7)	− 14 (0 8)			.057	497
(11)	*R*	.00053 (3.3) *3 5*	.15 (2 1)	− 13 (1 7)			− 12 (3 7) *−.82*	0066 (2 2)		− 036 (1 7)	.38 (3 3)			046	2386
Whites Only															
(8)	*lnWM*						.052 (4 2) *2*			042 (8 0)		035 (2 2)	−.012 (2 3)	24	2226
(8)	*lnWM*									035 (7 0)		078 (6 7)	−.027 (7 0)	23	
(9)	*N*					− 00017 (4 2) *− 45*	.26 (2 4) *29*	0057 (0.6)	− 064 (2 8)					020	2700
(9′)	*N*	− 00034 (1.9) *−.36*		−1 2 (1 5) *− 92*			33 (2 5) *37*	021 (1.0) *27*	− 071 (2 7) *.28*					020	
(9)	*N**					−.00015 (2 2) *− 42*	19 (1.0) *22*	.022 (0.5)	− 051 (1 2)					010	966
(10)	*Q*				00012 (0 7) *43*		15 (1 5) *17*			− 0053 (0 09)				021	356
(10′)	*Q*	0034 (2 0) *3 9*	97 (1 2) *1 1*				− 42 (1 3) *−0.5*			− 40 (1 8) *1 6*				.028	
(11)	*R*	.00056 (2 2) *4 2*	−.076 (0.7)	25 (1 6)			− 18 (3 3) *−1 4*	−.011 (1.9)		− 078 (2 1)				015	1978
(11)	*R**	00073 (1 6) *5 4*	.046 (0 5)	.53 (2 2)			− 20 (2 3) *−1 6*	− 028 (2 3)		− 10 (1 6)				014	697

Note Ratios of estimated regression coefficients to standard errors in parentheses, elasticities at means shown in italics

creases *WM* by 4.2 percent and a year of experience by 3.5 percent if the family has no children. Evidently caring for children reduces the intensity of training, and with three children, the effect of additional experience on mother's market earning power falls to zero. The results are somewhat sensitive to the exclusion of *WF* from equation (8), with the coefficient of *SM* falling by about one-sixth and those of *R′* and *N·R′* more than doubling. The number of children corresponding to a zero effect of experience on *WM* remains at close to three however.

The effect of an additional year of schooling on *WM*, the occupational wage measure, of 4.2 percent is smaller than that reported by either Mincer and Polachek or Steven Sandell and David Shapiro. An explanation of this difference is that selectivity bias has influenced the Mincer-Polachek and Sandell-Shapiro studies. In both studies wage data were available for only about 50 percent of the sample. Another possible explanation is that our average lifetime wage measure fails to reflect some wage gains attributable to experience.

The regression coefficients in equation (9) are largely as hypothesized. We are particularly interested in the coefficients of the price and income variables in the child quantity demand equation. Holding *Q* constant at its mean value of $2.30, a $1,000 increase in *WM* (34 percent) reduces the number of children by 0.4. This implies a partial elasticity of *N* with respect to *WM* of −.43 at the means (which is necessarily equal to the corresponding elasticity with respect to *Q*). Our results do not suggest that the "true" income effect on fertility is small relative to the effect of price. The

coefficient of *WF* is reasonably large relative to its standard error and implies an elasticity at the means of .29. This value is smaller than the "full income" elasticity, however, because father's earning power does not account for all of family full income. If father's earning power constituted two-thirds of full income, then the income and price elasticities implied by α_{25} and α_{26} would be about equal to each other. An increase in the husband's wage of $1.00 per hour (30 percent) is associated with an increase in the number of children of 0.26. An increase in husband's schooling, *ceteris paribus*, reduces the number of children. No pronounced partial correlation of the number of children with mother's age is observed, however.

In order to check the sensitivity of equation (9) to omission of the proxy for the price of child quality, we have estimated equation (9′). Using the elasticities calculated at the means, we compute the elasticity of N with respect to the price of child quality ($\pi_c N$) equal to .36, the elasticity with respect to the price of children ($\pi_c Q$) equal to −.59, and with respect to family income as represented by *WF* equal to .24. The $\bar{R}^2$ of equation (9′) is equal to that of equation (9). The implied own-price elasticity of N is about one-third larger in absolute value and the income elasticity about 20 percent smaller based on the unconstrained estimate of equation (9′). We conclude that the constrained estimate of equation (9) is not markedly worse than the unconstrained estimate.

In order to make sure the estimates of equation (9) pertain to *completed* fertility rather than to the desired timing of births, it has been reestimated for women who were 40–44 in 1967 (N^*).[6] The most striking result is the decline in the magnitude and significance of the *WF* coefficient, both absolutely and relative to that of *WM*. Evidently *WF* has a greater effect on the *timing* of births than on their ultimate number. This is consistent with the findings of Sue Ross. We do not compare our results quantitatively with the multiple equation models of Nerlove and Schultz or Cain and Dooley, because differences in model specification and variable definitions are substantial.

The results for equation (10) are disappointing. Only the coefficient of *WF* has the expected sign, and it alone can account for all of the variation in Q explained by the set of variables $WM \cdot N$, *WF*, and *SM*. None of the coefficients of the other variables approach statistical significance. Evidently, we have not been able to distinguish price from productivity effects in the Q equation. An additional problem is that offsprings' work experience, reflecting *OJT*, should probably be incorporated since Q is a market wage rate. Accounting for labor market experience in the model would seriously complicate the estimation procedure, however, and we have opted to ignore this variable in this study. All these reservations having been stated, it is still worth noting that the most significant regression coefficient, that of *WF*, is positive as hypothesized. Taking the coefficient of the price of child quality ($WM \cdot N$) to be zero, it follows that the substitution effect of price on the amount of child quality demanded is negative. The estimated true income elasticity of demand for child quality is no larger than that for child quantity.

Using a procedure like that for equations (9) and (9′), we have estimated equation (10′) to check sensitivity of the equation to the elimination of the proxy for price of number of children. The implied elasticity of Q with respect to the price of children is .76, with respect to the price of child quality, .29, and with respect to *WF*, −.11. The own-price elasticity remains positive while the income elasticity becomes negative—an even less satisfactory result than that obtained with the constrained estimate.

The estimated coefficients of *WM* and *WF* in equation (11) are qualitatively con-

[6]In order to eliminate possible cohort effects, we have also estimated equation (9) for the cohort of women who were 35–40 in 1967, comparing their number of children in 1967 (N_a^{**}) with the number in 1972 (N_b^{**}). The results, which suggest a weaker coefficient of *WF* relative to *WM* than do those where N is the dependent variable, are available from the authors on request.

sistent with the increase in the labor force participation of married women over time in that the elasticity of R with respect to WM is considerably larger than it is with respect to WF. Quantitatively, however, the results of equation (11) imply a much larger postwar increase in the labor force participation of married women than actually occurred.[7] Cain and Dooley report results similar to ours, using current labor force participation as the dependent variable. By contrast, Bowen and Finegan and Schultz conclude that the rising labor force participation of married women cannot be attributed to wage trends, as in their studies husband and wife wage effects cancel each other. Our results, while by no means definitive, lead us to believe that much can be gained from a multiple equation approach and a lifetime framework.

The most striking feature of the equation (11) estimate is the negligible influence of the number of children on the labor force participation of married women, in contrast with most previous studies which show a pronounced negative impact of children on labor force participation. It is noteworthy that Cain and Dooley's multiple equation model estimates also indicate a nonnegative impact of number of children on mother's labor force participation.[8] Two, not inconsistent, explanations of our results are: The true effect of N on R is indirect, running through the wage equation, as indicated by our results for equation (8); and we have not captured enough of the relevant variation in N with our instrument. In the underlying instrumental variable equation[9] for N the $\bar{R}^2$ is .10 for the entire sample, only .02 for whites, and .16 for blacks. The solution to this problem is not clear to us, as we have included a larger set of exogenous variables in our instrumental variable equations than most previous studies have had available.

[7]Rough calculations are available from the authors on request.

[8]Nerlove and Schultz, however, report a negative influence of the presence of *young* children on labor force participation, using Puerto Rican data.

[9]Available from the authors on request.

The positive coefficient of Q is also contrary to our expectations. Once again, the problem may lie in the difficulty in obtaining a good estimate of the Q instrumental variable equation; it may also involve correlation between Q and components of mother's market wage not captured by WM. The negative coefficient of age evidently reflects the dominance of a cohort effect over an aging effect, i.e., *ceteris paribus*, younger cohorts are more likely to participate in the labor force. The negative coefficient of SM presumably reflects the previously mentioned tendency of mothers with high levels of schooling to withdraw from the labor force during the years when young children are present. A more complicated functional form would be required to test the interpretation of this labor force withdrawal as attributable to the effect of schooling on productivity in child rearing.

III. Conclusion

Our empirical results encourage us to believe that a disaggregate multivariate approach is useful for the study of fertility and labor supply behavior. There is fairly persuasive evidence that the number of children demanded responds negatively to their cost and positively to family income, *ceteris paribus*. The true income elasticity of demand for children is evidently no smaller than that for child quality; it is evidently equal to or less than the own-price elasticity of demand for children.

Desired family size feeds back negatively to mother's market earning power. Students of male-female wage differentials have found that labor market experience increases women's earning power less than that of men. Our results suggest that declining family size will reduce this discrepancy in the future. Increased labor force attachment may prove to be a more powerful force toward male-female wage equality than "equal opportunity" labor market legislation.

The surprising lack of association between number of children and mothers' lifetime labor supply may be attributable to an inadequate instrument for N. However,

our result is consistent with the findings of Cain and Dooley, who used a somewhat differently specified multiple equation model and measured labor supply by current labor force participation.

On the negative side, our inability to represent the fertility variable with a suitable instrument is disappointing, and we have evidently been unsuccessful in controlling for efficiency in child quality production. Future research should investigate the nature of exogenous forces impinging on desired family size and variables which would enable us to distinguish efficiency in child rearing from labor market productivity.

Appendix—Variable Definitions

A detailed description of all variables is available from the authors on request. However, the following variables deserve particular attention.

WM—a measure of the wife's average lifetime market earning power. The measure used here is derived from two sources: the occupation of her longest job held between leaving school and birth of first child as recorded in the *NLS*; and the 1960 Census of Population data on median earnings of women in that occupation who worked 50–52 weeks in 1959, by race. This definition minimizes missing data. The selectivity bias due to using this measure is probably not nearly as severe as in the case of typical cross-section studies of wage rates on labor supply, since we have direct observation of this occupational wage for 80 percent of the *NLS* women. For the average *NLS* woman (for whom data are available) the "occupational wage" of the longest job since birth of first child is about equal to that of the longest job between leaving school and birth of first child.

Q—Measurement of children's market earning power was accomplished by merging data for the subset of women and their children who were all interviewed for the *NLS*s of Women, Young Men, and Young Women, respectively. Thus it becomes possible to use information supplied by the mother about her own characteristics and that of her parents as exogenous variables in developing instrumental variable equations for child quality. The measure of *Q* chosen is wage on current or last job for out-of-school youth. The regression coefficients of the instrumental variable equation for *Q* (in which the *child* is the unit of observation) are then used in conjunction with information on the exogenous variables for all (married, spouse present) women, whether or not they had children. Thus, a single child quality instrument is obtained for each woman for whom data are available on all the exogenous variables. When *Q* is a left-hand variable, the unit of observation becomes the child, rather than the mother (or, equivalently, the family).

WF—the father's (mother's husband) expected wage at age 40. This variable is itself an instrument, assumed to depend on all the other exogenous variables of the model; that is, father's average hourly rate of pay has been regressed on all the other exogenous variables. On the basis of these calculations, each family has been assigned an estimate of the father's wage at 40 years of age.

Additional exogenous variables are:

UR—a dummy variable equal to 1 if the mother lived in a rural area when she was 15 years of age.

P—a dummy variable equal to 1 if mother lived with both of her parents when she was 15 years of age.

J1—Duncan Index of the occupation of mother's father when mother was 15 years old.

J2—a dummy variable equal to 1 if mother's mother held a job outside the home when mother was 15 years old.

X—father's potential labor market experience, measured as his age minus 5 minus his number of years of schooling.

REFERENCES

G. S. Becker, "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, Universities-Nat. Bur. Econ. Res. conference series, Princeton 1960.

——— and **H. G. Lewis**, "On the Interaction Between the Quantity and Quality of Children," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, S279–88.

Y. Ben-Porath, "Economic Analysis of Fertility in Israel: Point and Counterpoint," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, S202–33.

William G. Bowen and T. Aldrich Finegan, *The Economics of Labor Force Participation*, Princeton 1969.

Glen Cain, *Labor Force Participation of Married Women*, Chicago 1966.

——— and **M. D. Dooley**, "Estimation of a Model of Labor Supply, Fertility, and Wages of Married Women," *J. Polit. Econ.*, Aug. 1976, *84*, S179–S201.

John F. Cogan, *Labor Supply and the Value of the Housewife's Time*, Santa Monica 1975.

M. Cohen, S. Rea, and R. Lerman, "A Micro-Model of Labor Supply," BLS staff paper no. 4, Washington 1970.

D. N. DeTray, "Child Quality and the Demand for Children," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, S70–95.

B. M. Fleisher, "Mother's Home Time and the Production of Child Quality," *Demography*, May 1977, *14*, 197–212.

R. Gronau, "Wage Comparisons—A Selectivity Bias," *J Polit. Econ.*, Nov./Dec., 1974, *82*, 1119–44.

J. J. Heckman, "Sample Selection Bias as a Specification Error," Rand Corp., Santa Monica 1976.

C. R. Hill and F. P. Stafford, "Allocation of Time to Preschool Children and Educational Opportunity," *J. Hum. Resources*, Summer 1974, *9*, 323–41.

H. S. Houthakker, "Compensated Changes in Quantities and Qualities Consumed," *Rev. Econ. Stud.*, Aug. 1952, *19*, 55–61.

A. S. Leibowitz, "Women's Allocation of Time to Market and Nonmarket Activities: Differences by Education," unpublished doctoral dissertation, Columbia Univ. 1972.

H. G. Lewis, "Comments on Selectivity Biases in Wage Comparisons," *J Polit. Econ.*, Nov./Dec. 1974, *82*, 1145–56.

R. T. Michael, "Education and the Desired Demand for Children," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, S128–64.

J. Mincer, "Labor Force Participation of Married Women," in H. Gregg Lewis, ed., *Aspects of Labor Economics*, Princeton 1962, 63–105.

———, "Market Prices, Opportunity Costs, and Income Effects," in Carl Christ, ed., *Measurement in Economics*, Stanford 1963.

——— and **S. Polachek**, "Family Investment in Human Capital: Earnings of Women," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, S76–S108.

M. Nerlove and T. P. Schultz, "Love and Life Between the Censuses: A Model of Family Decision Making in Puerto Rico, 1950–1960," Rand Corp., Santa Monica 1970.

R. Oaxaca, "Male-Female Wage Differentials in Urban Labor Markets," *Int. Econ. Rev.*, Oct. 1973, *14*, 693–709.

Sue G. Ross, *The Timing and Spacing of Births and Women's Labor Force Participation: An Economic Analysis*, New York 1974.

S. Sandell and D. Shaprio, "The Theory of Human Capital and the Earnings of Women: A Re-Examination of the Evidence," *J. Hum. Resources*, Winter 1978, *13*, 103–17.

T. Paul Schultz, *Estimating Labor Supply Functions for Married Women*, Santa Monica 1975.

H. Theil, "Qualities, Prices, and Budget Inquiries," *Rev. Econ. Stud.*, Oct. 1952, *19*, 129–47.

R. J. Willis, "A New Approach to the Economic Theory of Fertility Behavior," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, S14–64.

"National Longitudinal Surveys 1966–69," (*NLS*) Ohio State Univ.

U.S. Bureau of the Census, *U.S. Census of Population: 1960*, Subject Reports, *Occupational Characteristics*, Final Report PC(2)-7A, Washington 1973, Table 30.

Appropriative Water Rights and the Efficient Allocation of Resources

By H. STUART BURNESS AND JAMES P. QUIRK*

Historically, water rights to surface water in the United States have developed under two distinct legal doctrines—the English common law notion of riparian rights and the appropriative doctrine. Generally speaking, the riparian doctrine forms the basis for water law in the eastern states, while the western states have adopted the appropriative doctrine. Under the riparian doctrine, each property owner fronting on a lake or stream has a right to the unimpaired use of the waterway, regardless of the location of his property along the waterway and regardless of the time at which the property is acquired or use made of the waterway. Consequently, rights to water are only usufructuary: strictly speaking the right holder may not diminish the flow of water by physically consuming it as this would impair the rights of other riparians.

In practice the courts have held that "reasonable" diversions of water by riparian rights holders are permissible, but there are still severe restrictions on such diversions, coupled with uncertainty as to how a court will view any specific diversion. As a practical matter, the riparian doctrine is especially suited to an environment in which the use of water involves no diversions, for example, in the use of a stream for fishing, swimming, boating, transportation, or power generation.

In contrast, under the appropriative doctrine the right to a certain amount of water is established and maintained only through use; if there is a lapse in usage or a change in the nature of the usage, the right to the water can be lost.[1] Moreover, the right enables the holder to physically consume the water to which he is entitled, provided it is put to a beneficial use. Seniority of rights is based on the chronological order in which the right was obtained, the earliest user of water along a waterway being the most senior rights holder with priorities superior to those of junior rights holders. Under the appropriative doctrine, "first in time means first in right."

The appropriative doctrine was adopted in the West (and is spreading to eastern states as well) because it is well suited to the exploitation of a waterway under conditions in which the major uses of the waterway involve physical diversions of water, say for irrigation or for municipal or industrial uses. There are obvious advantages under such circumstances to a system of rights based on the appropriative doctrine, as discussed in the authors (1976); Charles Meyers; Jerome Milliman. In particular, an allocation of rights based on the appropriative doctrine preserves incentives for investment that would be foregone under the riparian scheme because of the common property characteristics of water under riparian allocation of rights.

This is not to say that the appropriative doctrine is without drawbacks from the point of view of economic efficiency. For example, inefficiencies can arise under the appropriative doctrine when an individual diverts more water than he can presently use profitably in order to establish a right to the use of such water in the future when the use might be profitable. To protect against this, state water laws limit appro-

*University of New Mexico and California Institute of Technology, respectively. This research was conducted at the Environmental Quality Laboratory at Caltech and was supported in part under a grant from the Energy Research and Development Administration, No. EX-76-G-03-1305, Caltech Energy Research Program.

[1] For a discussion of the legal principles involved under the riparian and appropriative doctrines, see Richard Dewsnut and Dallice Jensen.

priative rights to diversions that qualify as "beneficial consumptive use," thus excluding wasteful types or methods of water use. But there are obvious difficulties in establishing that water is being wasted by a rights holder, so that the protection afforded by the restriction of appropriations to beneficial consumptive use might be more illusory than real.[2]

Most of the allocative problems associated with the appropriative doctrine would be eliminated if water rights could be freely transferred or sold. But in every state operating under the appropriative doctrine, there are limitations on the transfer and sale of water rights. The statutes apply with most force to transfers that involve a change in use or in diversion location, as, for example, in the transfer of a water right from irrigation to industrial or power use, or in the transfer of water outside the property limits of the original rights holder.[3] Moreover, even when restrictions on intrastate transfers are relatively weak, sale or transfer of a water right that involves removal of water to another state is a practical impossibility, at least in the western states.

Independent of these statutory restrictions the appropriative doctrine provides an interesting scenario for the analysis of the efficiency aspects of water allocation. In this paper, we examine the efficiency implications of the appropriative doctrine at a long-run competitive equilibrium under simplified assumptions as to the legal status of water rights.[4] Briefly, our conclusions are the following. In the absence of a competitive market for the purchase and sale of water rights, the appropriative doctrine leads to an inefficient use of water. Inefficiency arises under the appropriative doctrine because of the unequal sharing of risks among the users of a waterway; senior appropriators bear less risk than junior appropriators. As an application of the Coase theorem, the introduction of a competitive market in water rights and use of diversion facilities eliminates allocative inefficiencies. However, increasing returns in the construction and maintenance of diversion facilities interferes with the establishment of competitive markets in the leasing of diversion capacity; furthermore, monopoly problems can arise in the market for water rights as well. Beyond this, limitations on entry can lead to problems associated with a suboptimal investment in diversion capacity.

For the special case of a waterway utilized by firms with identical production functions, allocative efficiency requires the equal sharing of risk (hence water) by all firms. But an assignment of water rights on the basis of equal sharing (a variant of the legal doctrine of correlative rights) leads to

[2]Meyers and A. Dan Tarlock cite a case in which use of water during the off-season to flood gophers from their holes was not deemed beneficial consumptive use. Furthermore, some court decisions have specified maximum amounts of water usage per irrigated acre that qualify as beneficial consumptive use. On the other hand, a large unnamed western irrigation district loses from 150,000 to 500,000 acre-feet of water yearly due to seepage in an unlined diversion canal, a method of use which could be considered wasteful. As this amount is included in its rights total, should it decide to line the canal, it could use the salvaged water.

[3]For example, in 1974, the Metropolitan Water District (*MWD*) of Southern California was able to transfer a portion of its rights to Colorado River water to the Southern California Edison Company, but only after the passage of enabling legislation by the California State Legislature, as Southern California Edison intended to use this water outside of the geographic limits of the *MWD*.

[4]For example, many existing statutes allow for revisions in priority in times of drought. As a consequence, junior domestic and municipal or industrial users might be satisfied prior to senior agricultural users although not without compensation. We ignore this complication in our analysis. The relevance of this point arises in conjunction with the recent western drought and the prediction of lower long-term water availability (for example, as suggested from the tree ring studies performed by the Lake Powell Research Group relative to the Colorado River). It is difficult to assess the importance of these matters, at least in the case of the Colorado River, as large accumulations of stored water and the Bureau of Reclamation's implicit policy of releasing enough water to satisfy all downstream users (which in total are limited to mean stream flow) suggest that it will be quite a while until such constraints become effective. We explore these questions to some extent in our 1977 paper.

much the same common property problems as the riparian scheme; and similar difficulties arise for the case of firms using diverse technologies. Thus, in the absence of freely transferrable property rights, the appropriative doctrine leads to an allocation of water that is inefficient, but alternative schemes for assigning water rights are generally not incentive compatible with a competitive environment.

I. The Model

The problems that arise for an efficient allocation of water under the appropriative doctrine rest ultimately on the random nature of water flows. We consider the case of a waterway with a flow of x acre-feet per year, where x is a random variable with known probability density function $f(x)$.[5] For simplicity, we ignore the autocorrelation of streamflows over time and concentrate instead on the characteristics of a steady-state situation. We assume that there are a number of potential users of the stream and that no institutional barriers to entry exist, except those associated with the rights of senior appropriators.

Under the appropriative doctrine, rights to water are established only through use. In order to use a_i units of water each period, the ith appropriator must have access to a diversion facility with a capacity at least equal to a_i units. In particular, firm i is assumed to possess a profit function $\pi^i(a_i, \bar{a}_i)$ where a_i is the use of water per period by firm i and $\bar{a}_i$ is the capacity of the diversion facility owned by firm i, subject to the restriction $a_i \leqq \bar{a}_i$. For simplicity we ignore other factors of production although clearly substitution of other factors for water could play an important role in the production process, particularly for firms with relatively junior rights.

We assume that there is no charge to an appropriator for the water he uses,[6] that $\pi_1^i \equiv \partial\pi^i/\partial a_i > 0$, $0 \leq a_i \leq \bar{a}_i$, $\pi_1^i = 0$ otherwise, and $\pi_{11}^i \equiv \partial^2\pi^i/\partial a_i^2 < 0$. Costs incurred in production are associated with the construction and maintenance of diversion facilities. It is clear that there are certain economies of scale associated with facilities such as pipelines and aqueducts. We assume that such nonconvexities apply for a certain range, after which problems of coordination, etc., overwhelm the natural economies of scale. In particular, we assume that $\pi_2^i \equiv \partial\pi^i/\partial\bar{a}_i < 0$, for $\bar{a}_i \geq 0$; $\pi_{22}^i \equiv \partial^2\pi^i/\partial\bar{a}_i^2 > 0$, for $\bar{a}_i < \bar{a}_i^*$; $\pi_{22}^i < 0$ for $\bar{a}_i > \bar{a}_i^*$. Moreover, in most of what follows we will assume that the profit-maximizing choices of diversion capacities occur in the range $\bar{a}_i > \bar{a}_i^*$, so that the marginal cost of adding diversion capacity is increasing. Finally, it is convenient to assume that diversion facilities deteriorate only through aging and not through use, so that $\pi_{12}^i \equiv \partial^2\pi^i/\partial a_i\partial\bar{a}_i = 0$. Under this assumption the profit function is separable in a_i and $\bar{a}_i$, so that $\pi^i(a_i, \bar{a}_i) = R^i(a_i) - C^i(\bar{a}_i)$, where R^i and C^i are the revenue and cost functions for the ith firm.

Our primary purpose is to identify the sources of allocative inefficiency associated with the appropriative doctrine. These sources are most easily identified in the simplest possible setting. Hence our approach in the body of this paper is to examine in detail the special case where all appropriators have identical profit functions, with each appropriator acting to maximize expected profits. Extensions of these results to cases of dissimilar or risk-averse firms are noted when of interest.

With this as background we examine the long-run equilibrium of a waterway being exploited under the system of appropriative rights. We label rights holders in order of seniority, with firm 1 being the most senior rights holder, firm 2 second in seniority, etc. Clearly, in long-run equilibrium with

[5]We assume $f(x) \geq 0$ for $x \geq 0$, $f(x) = 0$ for $x < 0$. Letting

$$F(x) = \int_0^x f(c)dc$$

we have $F(0) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

[6]Generally, the presence of charges do not affect the results, hence the simpler formulation; note that delivery charges are incorporated into the costs of constructing and maintaining diversion facilities.

known probability density function $f(x)$, no expected profit-maximizing firm would acquire a diversion facility with capacity in excess of its rights to use water; moreover, rights in excess of diversion capacity would not be approved by the state rights administrator. Hence $\bar{a}_i$ can be identified as the appropriative rights of firm i. As a matter of notation, let

$$A_i = \sum_{j=1}^{i} \bar{a}_j$$

Then A_i denotes the aggregate amounts of claims to water senior to the claims of firm $i + 1$; alternatively, A_i is the total amount of diversion capacity owned (or leased) by firms 1 through i ($A_0 = 0$).

Identifying water rights with diversion capacities, the assignment of rights under the appropriative doctrine can be summarized in the vector $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_N)$, where there are N firms exploiting a stream. Expected profits for firm i, $E^i\pi$, are then given by

$$E^i\pi = F(A_{i-1})\pi(0,\bar{a}_i) + \int_{A_{i-1}}^{A_i} \pi(x - A_{i-1},\bar{a}_i) f(x)\,dx + [1 - F(A_i)]\pi(\bar{a}_i,\bar{a}_i)$$

where $\pi^i = \pi$ for $i = i, \ldots, N$ (all firms are identical) and $F(c) = \int_0^c f(x)dx$. Thus, firm i receives zero units of water if the streamflow x is no more than enough to satisfy senior claimants; the probability that river flows do not exceed A_{i-1} is $F(A_{i-1})$ while profits for the ith firm in this case are $\pi(0,\bar{a}_i)$, so that the expected value of this outcome is represented by the first term in the expression for expected profits. If the flow exceeds senior claims and can be handled by firm i's diversion capacity, then expected profits are given by the second term in this expression; that is, expected profits are $\pi(x - A_{i-1},\bar{a}_i)$ times $f(x)$ summed over the interval of river flows which yield increasing amounts of water to the ith firm. If river flow exceeds the capacity of claimants 1 through i, then the ith firm receives its entire appropriation. The probability of this occurrence is $1 - F(A_i)$ and ith firm profits are $\pi(\bar{a}_i, \bar{a}_i)$, hence the third term in the expression.

II. Water Rights and Water Usage: Appropriative System

Clearly, senior claimants obtain a preferred position due to their priority in access to the streamflow. Let x_i denote the flow available for use by firm i and let $G^i(x_i)$ denote the cumulative probability distribution over this flow. Then $G^i(x_i)$ is given by

$$G^i(0) = F(A_{i-1})$$
$$G^i(x - A_{i-1}) = F(x), \quad A_{i-1} \leq x \leq \infty$$

Since $A_j > A_i$ for $j > i$, it follows that $G^i(b) \leqq G^j(b)$ for $b \geqq 0$, $j > i$, with strict inequality for $b \geq A_{i-1}$, assuming $f(x) > 0$ for $A_{i-1} \leq x \leq A_i$.

Hence the probability distribution of streamflows facing a junior appropriator is *stochastically dominated* (in the sense of first-degree stochastic dominance) by the distribution facing any senior appropriator. Under the assumption of positive marginal profitability of water use ($\pi_1(a_k) > 0$ for $a_k \geq 0$, $k = i, j$) stochastic dominance implies that for any monotonically increasing measurable utility function U over profits,

$$E_{G^i} U(\pi) > E_{G^j} U(\pi)$$

for $i < j$ (see Quirk and Rubin Saposnik or Josef Hadar and W. R. Russell). This result can be summarized as follows:

PROPOSITION 1: *Under the appropriative doctrine, the probability distribution facing any senior appropriator is unambiguously preferred by any potential user of the waterway to that facing a junior appropriator.*

To analyze the consequences of the appropriative doctrine for the allocation of water among potential users, we consider the distribution of rights that would arise under stationary conditions, assuming free entry coupled with an absolute prohibition on the sale or transfer of water rights to other water users or alternative uses. Given that senior claims $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_{i-1})$ exist, firm i chooses the diversion capacity $\bar{a}_i$ which maximizes the expected utility of

profits. Assuming that firm i is risk neutral, expected utility maximization implies expected profit maximization where

$$E^i\pi = F(A_{i-1})\pi(0,\bar{a}_i) + \int_{A_{i-1}}^{A_i} \pi(x - A_{i-1},\bar{a}_i)\, f(x)dx + [1 - F(A_i)]\pi(\bar{a}_i,\bar{a}_i)$$

For $\bar{a}_i > 0$ we have

$$\frac{dE^i\pi}{d\bar{a}_i} = F(A_{i-1})\pi_2(0,\bar{a}_i) + \int_{A_{i-1}}^{A_i} \pi_2(x - A_{i-1},\bar{a}_i)\, f(x)dx + [1 - F(A_i)\{\pi_1(\bar{a}_i,\bar{a}_i) + \pi_2(\bar{a}_i,\bar{a}_i)\} = 0$$

Under separability of the profit function ($\pi^i_{12} = 0$) we can write $\pi_1(z,w) = \pi_1(z)$ and $\pi_2(z,w) = \pi_2(w)$ and the first-order condition reduces to

$$\pi_2(\bar{a}_i) + [1 - F(A_i)]\pi_1(\bar{a}_i) = 0$$

which we write as

$$M_i(\bar{a}_i) = \pi_2(\bar{a}_i) + [1 - F(A_{i-1} + \bar{a}_i)]\pi_1(\bar{a}_i) = 0$$

At a regular maximum of expected profits we have $\partial M_i(\bar{a}_i)/\partial \bar{a}_i < 0$. Observing that

$$\left(\frac{\partial M_i(\bar{a}_i)}{\partial A_{i-1}}\right)_{\bar{a}_i \text{constant}} = -\pi_1(\bar{a}_i)f(A_i) < 0$$

we have

PROPOSITION 2: *Given two expected profit-maximizing firms with identical separable profit functions, the firm with senior rights claims a larger quantity of water (constructs a larger diversion capacity) than does a firm with junior rights.*

The incentive rationale underlying Proposition 2 is obvious since senior firms face preferred probability distributions over streamflows relative to junior firms. Already there is some indication of the allocative inefficiency of the appropriative system in the absence of a competitive market for water rights. At an optimum, firms with identical production and profit functions should presumably divert and use identical amounts of water. The appropriative system biases the distribution of water use in favor of firms with earlier filing dates for water rights over firms filing later in time.

Proposition 2 generalizes directly when firms are risk-averse expected utility maximizers. If firms face different technologies, then the ordering of diversion capacities does not generalize. However, the intent of the proposition does. Fundamentally, the first-order conditions imply that with all firms operating in competitive output markets, junior appropriators are more productive at the margin in the sense that the ratio of marginal revenue to marginal cost increases with decreasing seniority. If firms are identical, this implies that senior firms appropriate more water.

An issue of some importance to allocative efficiency of the appropriative doctrine is the extent to which the flow of a waterway is appropriated. To put it in other terms, how much diversion capacity is built under the system of appropriative rights, assuming each firm builds its own capacity?

Let N denote the number of appropriators exploiting a waterway at a long-run competitive equilibrium so that the last firm just finds it worthwhile to appropriate a portion of the stream by building a diversion capacity. Assuming risk neutrality and separable profit functions for all firms, if the waterway is appropriated by N firms we have $E^N\pi \geq 0$ for $\bar{a}_N > 0$, $E^{N+1}\pi < 0$ for $\bar{a}_{N+1} > 0$, where

$$\pi_2(\bar{a}_N) + [1 - F(A_N)\pi_1(\bar{a}_N) = 0$$

$$\pi_{22}(\bar{a}_N) - f(A_N)\pi_1(\bar{a}_N) + [1 - F(A_N)]\pi_{11}(\bar{a}_N) \leq 0$$

From the first-order condition we have

$$F(A_N) = 1 + \frac{\pi_2(\bar{a}_N)}{\pi_1(\bar{a}_N)}$$

where $\pi_2(\bar{a}_N) < 0$, $\pi_1(\bar{a}_N) > 0$ for all $\bar{a}_N \geq 0$. The entire stream is completely appropriated only if $\bar{a}_N \to 0$ with

$$\lim_{\bar{a}_N \to 0} \pi_1(\bar{a}_N) = +\infty$$

So long as π_1 is bounded from above, the total amount of diversionary capacity

built is less than the maximum flow of the stream.

Given a neoclassical production function and a competitive output market for the firm's product, then

$$\lim_{\bar{a}_N \to 0} \pi_1(\bar{a}_N) = +\infty$$

This does not necessarily guarantee a completely appropriated stream, however. In fact, under extreme conditions of scale economies in diversion capacity

$$\lim_{\bar{a}_N \to 0} \pi_2(\bar{a}_N) = -\infty$$

We will assume that increasing returns "dominate" for small diversion capacities in the sense that

$$\lim_{\bar{a}_i \to 0} \left(\frac{-\pi_2(\bar{a}_i)}{\pi_1(\bar{a}_i)}\right) > 1$$

Under such circumstances, the number of firms exploiting the stream is finite, with each firm of noninfinitesimal size.

PROPOSITION 3: *If increasing returns dominate for small diversion capacities and if the potential users of a stream are risk neutral with identical separable profit functions, then the aggregate amount of water rights (diversion capacity) at a long-run competitive equilibrium is less than the maximum flow of the stream; further, each appropriator is of noninfinitesimal size, and the number of appropriators is finite.*[7]

COROLLARY: *Under the conditions of Proposition 3, the expected value of streamflows exceeds the expected value of diversions.*

III. Allocative Inefficiency of the Appropriative System

Consider next a waterway operating in long-run competitive equilibrium, exploited by N risk-neutral firms with identical separable profit functions, with water rights determined under the appropriative doctrine. Suppose the conditions of Proposition 3 hold, so that N is finite. Let $(\bar{a}_1, \ldots, \bar{a}_N)$ denote the vector of diversion capacities for the N firms. Associated with this pattern of appropriations is a value of aggregate expected profits E^A, given by

$$E^A = \sum_{i=1}^{N} \{F(A_{i-1})\pi(0, \bar{a}_i) + \int_{A_{i-1}}^{A_i} \pi(x - A_{i-1}, \bar{a}_i) f(x)dx + [1 - F(A_i)]\pi(\bar{a}_i, \bar{a}_i)\}$$

Given A_N the diversion capacity under the appropriative system, is the pattern of investment in diversion capacity and use of water associated with the appropriative system efficient? The answer is that the appropriative system is not efficient. We establish this by showing that there exists a feasible alternative to the appropriative system, namely equal sharing, that produces a higher value of output for every possible streamflow.

Given the aggregate diversion capacity A_N and given any streamflow x, then aggregate profits associated with an arbitrary feasible allocation of diversion capacities and water usage are given by

$$\sum_{i=1}^{N} \pi(a_i, \bar{a}_i) \quad \text{where } a_i \leqq \bar{a}_i,$$
$$i = 1, \ldots, N,$$
$$\sum_{i=1}^{N} \alpha_i \leqq x, \qquad \sum_{i=1}^{N} \bar{a}_i = A_N$$

It is immediate that with $\pi_{11} < 0$, then so long as $\pi_{22} < 0$ (marginal cost of adding diversion capacity is increasing), aggregate profits are maximized with $a_i = x/N$ for $x \leq A_N$, $a_i = A_N/N$ for $x > A_N$, with $\bar{a}_i = A_N/N$ for $i = 1, \ldots, N$. In fact, writing aggregate profits as

$$\sum_{i=1}^{N} \pi(a_i, \bar{a}_i) = \sum_{i=1}^{N} \{R(a_i) - C(\bar{a}_i)\}$$

[7] Proposition 3 generalizes even for nonidentical risk-averse expected utility maximizers facing diverse technologies. To see this observe that the last appropriator must be of noninfinitesimal size else the dominance of increasing returns for small diversion capacity is violated. Given this, complete appropriation of the river implies that the costs of constructing diversion capacity be negative at some level, a clear impossibility (see Appendix A).

it also follows that under equal sharing, aggregate revenue $\Sigma_{i=1}^N R_i(a_i)$ is maximized and aggregate cost $\Sigma_{i=1}^N C(\bar{a}_i)$ is minimized, subject to the feasibility constraints. Finally, for $N > 1$, aggregate profits (and aggregate revenue) are clearly less under the appropriate system than under equal sharing, for any value of x, because of the ordering of capacities from Proposition 2. Assuming competitive input and output markets, the results in turn imply that Pareto optimality implies equal sharing, while the appropriative system is inefficient.

PROPOSITION 4: *Assume that N firms exploit a waterway, with each firm having an identical separable profit function strictly concave in water usage. If the marginal cost of adding diversion capacity is increasing, then equal sharing is the efficient allocation of diversion capacity and water usage; allocation under the appropriative system is inefficient.*[8]

The assumption that the marginal cost of adding diversion capacity is increasing at an equal sharing allocation is restrictive. As it turns out, when this assumption is relaxed to permit declining marginal costs of adding diversion capacity, then it can still be shown that the appropriative system is inefficient in the sense that expected profits under other feasible allocations exceed those under the appropriative system. However, equal sharing is no longer necessarily the efficient allocation. Details are given in Appendix B.

The source of the allocative inefficiencies of the appropriative system is unequal sharing of risk and diversion capacity among firms. The Coase theorem makes it clear that a solution to the problem involves the establishment of competitive markets in water rights. Let α_{ij} be the fraction of firm i's rights that is purchased by firm j, and let β_{ij} be the fraction of firm i's diversion capacity leased to firm j; let p_i be the price for a 1 percent share of firm i's water and let q_i be the price for a 1 percent share of firm i's diversion capacity. Then given the investment vector $(\bar{a}_1, \ldots, \bar{a}_N)$ established under the appropriative allocation[9] and assuming risk neutrality, with competitive markets in rights and capacity, each firm picks α_{ij} and β_{ij} so as to maximize expected profits. At an equilibrium (see Appendix C for a complete derivation) with $\bar{a} = A_N/N$, we have

$$(1)\quad \int_{A_{i-1}}^{A_i} \pi_1\left(\frac{x}{N}\right)(x - A_{i-1})\, f(x)dx + \sum_{r=i+1}^{N} \int_{A_{r-1}}^{A_r} \pi_1\left(\frac{x}{N}\right)\bar{a}_i\, f(x)dx + [1 - F(A_N)]\pi_1(\bar{a})\bar{a}_i = p_i + q_i, \quad i = 1, \ldots, N$$

$$(2)\quad q_j/\bar{a}_j = q_i/\bar{a}_i \qquad i,j = 1, \ldots, N$$

Condition (1) can be written as

$$(1')\quad \int_{A_{i-1}}^{A_i} \pi_1\left(\frac{x}{N}\right)\left(\frac{x - A_{i-1}}{\bar{a}_i}\right) f(x)dx + \int_{A_i}^{A_N} \pi_1\left(\frac{x}{N}\right) f(x)dx + [1 - F(A_N)]\pi_1(\bar{a}) = p_i/\bar{a}_i + q_i/\bar{a}_i$$

In (2), $q_i > q_j$ for $i < j$; the price for 1 percent of a senior firm's capacity exceeds that of a junior firm, as the senior firm has larger capacity; however, $q_i/\bar{a}_i = q_j/\bar{a}_j$ for $i, j = 1, \ldots, N$, so that price per unit of capacity is equal among all firms. Since p_i is the price for 1 percent of firm i's water rights, 100 x $p_i/\bar{a}_i$ is the price of obtaining one unit of firm i's water when available by streamflow. Then the left-hand side of (1′) (multiplied by 100) is the expected marginal

[8]Proposition 4 generalizes as well. Note, however, that with diverse technologies, at the optimum water is prorated among firms according to their productivity so the expected marginal profitability is zero across firms, a condition that implies equal sharing if firms are identical.

[9]Should the eventuality of resale of water rights be foreseen, one might question the determinancy at the investment vector $(\bar{a}_1, \ldots, \bar{a}_N)$: what prevents a senior appropriator from "over-appropriating" for possible future resale? Fortunately this poses no problem as the appropriative doctrine is clear on this matter: to obtain a right to water it must be diverted, and diversions are limited to beneficial consumptive use. However in practice this may be problematic (see fn. 2).

profitability of a unit of water obtained from firm i, set equal to its marginal cost, including the increase in cost due to the added diversion capacity necessary to deliver the water. It is clear from (1) and (1′) that $i < j$ implies $p_i > p_j$ and $p_i/\bar{a}_i > p_j/\bar{a}_j$, but at the equilibrium prices p_i, $i = 1, \ldots, N$, purchasers of water are indifferent among suppliers at the margin.

Clearly conditions (1) and (2) are consistent with market clearing. Hence an efficient mix of capacities and usages (equal sharing) can be attained under the appropriative system, given competitive markets in water rights and diversion capacity and given a fixed diversion capacity. Thus we have

PROPOSITION 5: *Given that all firms have identical separable profit functions and are risk neutral, and given that increasing returns dominate for small diversion capacities, the appropriative system of rights allocation coupled with competitive markets in rights and diversion facilities leads to an efficient outcome (namely, equal sharing), given the fixed aggregate installed diversionary capacity. The price per unit of water varies monotonically with the seniority of the water supplier; the price per unit of capacity is constant across firms.*[10]

However, there are some problems with the conclusion of Proposition 5. Economies of scale in the delivery system for water are pervasive enough that it is difficult to maintain a competitive environment in the market for diversionary facilities and hence in the market for water rights as well. In fact, it is this phenomeon that no doubt accounts for the creation of publicly controlled irrigation districts in the Southwest, designed to achieve the savings from scale while minimizing monopolistic distortions; and this also accounts for the (rarely enforced) acreage limitations on recipients of water from the Bureau of Reclamation projects. Admittedly, in principle monopoly distortions could also be eliminated through appropriate bribes, but excessive transactions costs impose impediments to such a policy.

Proposition 5 takes as given the aggregate diversion capacity that is built under the appropriative system and asserts that competitive markets in rights and in leases of diversion facilities lead to an efficient outcome given that diversion capacity. This still leaves unanswered the issue of an optimal level of aggregate diversion capacity for a waterway.

We begin by examining a variant of this problem. Suppose that each firm owns its own diversion facility, and that increasing returns dominate for small diversion capacities, with N firms exploiting a waterway in long-run equilibrium under the appropriative doctrine. Aggregate capacity is A_N units. Consider in contrast the same N firms (all with identical separable profit functions) operating under equal sharing of water, with identical diversion capacities. What can be said about the amount of capacity that will be installed under equal sharing if aggregate expected profits are to be maximized?

First-order conditions require that

$$\pi_2(\bar{a}_N) + [1 - F(A_N)]\pi_1(\bar{a}_N) = 0$$

for the appropriative scheme, while

$$\pi_2(\bar{a}) + [1 - F(N\bar{a})]\pi_1(\bar{a}) = 0$$

for the equal sharing scheme. Thus,

$$F(A_N) = 1 + \frac{\pi_2(\bar{a}_N)}{\pi_1(\bar{a}_N)}$$

and

$$F(N\bar{a}) = 1 + \frac{\pi_2(\bar{a})}{\pi_1(\bar{a})}$$

$$F(A_N) - F(N\bar{a}) = \frac{\pi_2(\bar{a}_N)}{\pi_1(\bar{a}_N)} - \frac{\pi_2(\bar{a})}{\pi_1(\bar{a})}$$

We have $\pi_1 > 0$ for $a_i \geq 0$, $\pi_2 < 0$ for $\bar{a}_i \geq 0$. Further, assume that all firms operate under nondecreasing marginal costs of diversion capacity; in particular, $\bar{a} \geq \bar{a}_N$ implies $\pi_2(\bar{a}) \leq \pi_2(\bar{a}_N)$.

Suppose that $N\bar{a} \geq A_N$. Then $\pi_2(\bar{a}_N)/\pi_1(\bar{a}_N) - \pi_2(\bar{a})/\pi_1(\bar{a}) \leq 0$. But $N\bar{a} \geq A_N$ implies $\bar{a} > \bar{a}_N$ for $N > 1$. Hence $0 < \pi_1(\bar{a}) <$

[10]Proposition 5 generalizes in the same manner as Proposition 4.

$\pi_1(\bar{a}_N)$ and $0 > \pi_2(\bar{a}_N) > \pi_2(\bar{a})$ so that $\pi_2(\bar{a}_N)/\pi_1(\bar{a}_N) - \pi_2(\bar{a})/\pi_1(\bar{a}) > 0$, a contradiction. Hence $N\bar{a} < A_N$.

PROPOSITION 6: *Let N be the number of firms with identical separable profit functions exploiting a waterway in long-run equilibrium under the appropriative system. Then the aggregate diversion capacity constructed by these N firms exceeds that which would be constructed by the same firms under equal sharing, assuming that each firm builds its own diversion facility and that the marginal cost of constructing diversion facilities is increasing.*[11]

Thus there is a systematic overinvestment in diversion capacity under the appropriative scheme, assuming that the same N firms exploit a waterway under either equal sharing or the appropriative scheme, with each firm building its own diversion capacity.

IV. A Diversion Capacity Industry

It is clear, however, that with increasing returns operating with respect to diversion facilities, one would expect the development of an independent subindustry engaged in the construction and leasing of such facilities. As noted earlier, there are monopolistic problems present in such an industry; this has led to the formation of publicly operated and controlled irrigation districts which act in effect as lessors of diversion capacity. Suppose that the monopoly problems are overcome so that there is competitive pricing of leases. Let $C(\bar{a})$ denote the annualized cost associated with a diversion facility of capacity $\bar{a}$. Then under long-run competitive conditions, the aggregate diversion capacity for a waterway would be equal to $M\bar{a}^*$, where M is the number of leasing firms and $\bar{a}^*$ is the capacity owned by any one leasing firm, with $C(\bar{a}^*)/\bar{a}^* = C'(\bar{a}^*)$. That is, each leasing firm builds a capacity such that the average annualized cost per unit of capacity is minimized. Under competitive conditions, the charge for leasing a unit of capacity would then equal $C'(\bar{a}^*)$. Thus lessees would face constant marginal costs of diversion facilities; i.e., $\pi_2(\bar{a}) = -C'(\bar{a}^*)$ is a constant independent of $\bar{a}$.

Recall that under the appropriative scheme,

$$\pi_2(\bar{a}_N) + [1 - F(A_N)]\pi_1(\bar{a}_N) = 0$$

With a leasing industry operating under competitive conditions, we have $\pi_2(\bar{a}_N)$ independent of $\bar{a}_N$. Given that $C'(\bar{a}^*)(\equiv -\pi_2(\bar{a}))$ is less than $\lim_{a\to 0}\pi_1(a,a)$, it is clear that the "marginal" firm chooses a capacity that approaches zero as $N \to +\infty$. (If $C'(\bar{a}^*) > \lim_{a\to 0}\pi_1(a)$, then no firm finds it profitable to exploit the waterway). It follows that

$$F(A_N) = 1 - \frac{C'(\bar{a}^*)}{\lim_{a\to 0}\pi_1(a)}$$

The same condition holds when aggregate expected profits are maximized, with all firms sharing equally in streamflows, since at a maximum of aggregate expected profits we have

$$\pi_2(\bar{a}) + [1 - F(N\bar{a})]\pi_1(\bar{a}) = 0$$

so that

$$F(N\bar{a}) = 1 - \frac{C'(\bar{a}^*)}{\lim_{a\to 0}\pi_1(a)}$$

Hence $A_N = Na$. We summarize this as

PROPOSITION 7: *Suppose there is an arbitrary number of firms, all with identical separable profit functions, exploiting a waterway. In addition suppose there is a competitive industry in diversion capacity which leases capacity to rights holders, with each leasing firm building the capacity which minimizes the average annualized cost per unit of capacity. If entry is free and unobstructed in both the diversion capacity and water-using industries, then in long-run equilibrium aggregate investment is the same*

[11] Proposition 6 relies heavily on separability of the profit function. Although the proof is less direct the result generalizes for diverse technologies (maintaining the separability assumption) but not for risk averters.

under either the appropriative or equal sharing systems.[12]

Hence, the establishment of a competitive leasing industry that takes full advantage of the economics of scale in diversion facilities leads to the same aggregate diversion capacities under the appropriative scheme as under equal sharing.[13]

From Proposition 4 we know that for any given diversion capacity, equal sharing is the efficient solution given competitive markets with aggregate profits and revenue larger for any value of stream flows x than under the appropriative system. Thus with a competitive leasing industry operating to capture the limited economies of scale in building diversion capacity, equal sharing is a necessary condition for Pareto optimality, and involves the same aggregate diversion capacity as the appropriative system. Finally, the equal sharing allocation can be achieved under an appropriative system by competitive markets in water rights, from Proposition 5. By employing l'Hospital's rule we have

$$\lim_{N \to \infty} NR\left(\frac{a}{N}\right) = aR_1(0)$$

so that the aggregate revenue function for equal sharing is continuous at the origin. By a limiting argument we have

PROPOSITION 8: *Under the conditions of Proposition 7, at a long-run equilibrium equal sharing is a necessary condition for Pareto optimality and equal sharing can be achieved under an appropriative system through competitive markets in water rights.*

The appropriative system possesses one fundamental advantage over the riparian system or the equal sharing of rights in that it provides *tenure certainty* for each rights holder: rights holders are in principle protected against loss of their rights through the legal actions of others. While the appearance of a new claimant to water can dilute the privileges of existing water users under either the riparian or equal sharing systems, the principle of "first in time means first in right" protects the privileges of existing users under the appropriative system.

Unfortunately in practice tenure certainty is difficult to guarantee even under the appropriative system. Due to spatial dispersion of appropriators, informational inadequacies and random elements (such as variability in return flows), it is often difficult to determine whether a diminished downstream flow to senior appropriators is the result of the stochastic nature of river flows or the improper actions of upstream junior appropriators. And, as we have seen, the principle of tenure certainty is bought at the cost of economic efficiency, so long as water rights are not freely transferable.

Limitations on the transferability of water exist in the form of federal and state statutes, and interregional and interstate compacts. Moreover, there are other impediments to transfers: fixed diversion capacities, transactions costs, and externalities. Externalities arise because a change in the nature or location of water diversions can affect return flows to a river and hence can impinge on the established water rights of third parties. Thus there are sound economic grounds for certain of the existing limitations on rights transfers.

However, we would argue that considerable potential latitude for the transfer of water rights still exists, and that economic efficiency could be improved by weakening existing legal constraints on such transfers. The usual argument in favor of transferable water rights identifies the higher productivity of water in industrial or municipal use as compared to present usage which is highly concentrated in irrigated farming. While we certainly agree with this argument, our conclusion goes even further: even among identical firms producing identical products, freely transferable water rights leads to increased economic efficiency.

[12]This result generalizes directly for diverse technologies and risk-averse expected utility maximizers.

[13]Although Proposition 7 is instructive, one would not expect it to be operational in the real world because of spatial monopolies in the diversion leasing industry.

Our approach in this paper has centered on the simple case of a static long-run competitive equilibrium with an uncontrolled river. We have not attempted to model the dynamics of the process by which rights are acquired and implemented under the appropriative system, nor have we examined the special problems that arise when a reservoir system is constructed with releases to downstream users being determined in an optimal fashion, subject to the priorities that hold under the appropriative system of rights.[14] It is clear to us, however, that whatever are the complexities introduced into the analysis by these factors, there are still advantages that can be gained by widening the possibilities for transferability of water rights.

APPENDIX A

Proposition 3 generalizes immediately for nonidentical firms as no comparisons are made across firms. The result generalizes as well for risk-averse producers—in fact even for nonidentical firms with nonidentical utility functions, given monotone preferences and risk aversion.

To see this first observe that first-order conditions for expected utility maximization require that

$$\frac{E^i u(\pi^i)}{d\bar{a}_i} = 0 = F(A_{i-1})u'[\pi^i(0,\bar{a}_i)]\pi_2^i(\bar{a}_i) + \int_{A_{i-1}}^{A_i} u'[\pi^i(x - A_{i-1},\bar{a}_i)]\pi_2^i(\bar{a}_i)f(x)dx + [1 - F(A_i)]u'[\pi^i(\bar{a}_i,\bar{a}_i)] \cdot [\pi_1^i(\bar{a}_i) + \pi_2^i(\bar{a}_i)]$$

subject to $E^i u(\pi) \geq u(0)$. Suppose the last appropriator chooses capacity $\bar{a}_N = 0$, satisfying the first-order condition as an equality. Then for any u, monotonic and concave, the first-order condition becomes

$$\lim_{\bar{a}_N \to 0} \{\pi_2^N(\bar{a}_N) + [1 - F(A_N)]\pi_1^N(\bar{a}_N)\} = 0$$

This conflicts with the assumption that increasing returns dominates for small diversion capacities. Hence $\bar{a}_N$ is noninfinitesimal and N is finite. If the entire river is appropriated these first-order conditions become

$$F(A_{N-1})u'[\pi^N(0,\bar{a}_N)]\pi_2^N(\bar{a}_N) + \int_{A_{N-1}}^{A_N} u'[\pi^N(x - A_{N-1},\bar{a}_N)] \pi_2^N(\bar{a}_N)f(x)dx = 0$$

which is impossible in view of the negativity of π_2. Hence the river is not fully appropriated.

APPENDIX B

Consider an arbitrary reassignment of diversion capacities and water rights among the N firms such that firm j receives β_j percent of A_N as its diversion capacity along with α_{ij} percent of any streamflow within the range A_{i-1} to A_i where

$$\beta_j \geq 0, \sum_{j=1}^{N} \beta_j = 1, \alpha_{ij} \geq 0, \qquad \sum_{j=1}^{N} \alpha_{ij} = 1, i, j = 1, \ldots, N$$

Let $\alpha = [\alpha_{ij}], \beta = [\beta_j]$ and let $E(\alpha,\beta)$ be the expected value of aggregate profits, where

$$E(\alpha,\beta) = \sum_{j=1}^{N} \{\sum_{i=1}^{N} \int_{A_{i-1}}^{A_i} \pi[\alpha_{ij}(x - A_{i-1}) + \sum_{k=1}^{i-1} \alpha_{kj}\bar{a}_k, \beta_j A_N] f(x)dx + [1 - F(A_N)]\pi(\beta_j A_N, \beta_j A_N)\}$$

Thus $E(\alpha\beta) = E(\bar{\alpha},\bar{\beta}) = E^A$ for $\bar{\beta}_j = \bar{a}_j/A_N$, $\bar{\alpha}_{jj} = 1, \bar{\alpha}_{ij} = 0, i \neq j, \quad i,j = 1,\ldots,N$

$$\text{Let } L = E(\alpha,\beta) + \lambda(1 - \sum_{j=1}^{N} \beta_j) + \sum_{i=1}^{N} \mu_i(1 - \sum_{j=1}^{N} \alpha_{ij})$$

At a constrained maximum of $E(\alpha,\beta)$ we have

$$\frac{\partial E(\alpha,\beta)}{\partial \beta_r} - \lambda \leq 0 \quad r = 1,\ldots,N$$

$$(<0 \text{ implies } \beta_r = 0)$$

[14]See the authors (1977) for a simplified treatment of the problem of reservoir management under the appropriative system.

$$\frac{\partial E(\alpha,\beta)}{\partial \alpha_{sr}} - \mu_r \leq 0 \quad s, r = 1,\ldots,N$$

$$(<0 \text{ implies } \alpha_{sr} = 0)$$

where, given $\pi_{12} = 0$,

$$\frac{\partial E(\alpha,\beta)}{\partial \beta_r} = A_N\{\pi_2(\beta,A_N) + [1 - F(A_N)]\pi_1(\beta,A_N)\} \quad r = 1,\ldots,N$$

$$\frac{\partial E(\alpha,\beta)}{\partial \alpha_{sr}} = \int_{A_{s-1}}^{A_s} \pi_1[\alpha_{sr}(x - A_{s-1}) + \sum_{k=1}^{s-1} \alpha_{kr}\bar{a}_k](x - A_{s-1})f(x)dx + \sum_{t=r+1}^{N} \int_{A_{t-1}}^{A_t} \pi_1[\alpha_{tr}(x - A_{t-1}) + \sum_{k=1}^{t-1} \alpha_{kr}\bar{a}_k]\bar{a}_s f(x)dx$$

$$r,s = 1,\ldots,N$$

The question we pose is whether the appropriative system $(\bar{\alpha},\bar{\beta})$ qualifies as a candidate for a maximizer of $E(\alpha,\beta)$. Evaluate the expression immediately above at $\alpha_{ss} = 1$, $\alpha_{sr} = 0, s \neq r, \beta_r = \bar{a}_r/A_N, r, s = 1,\ldots,N$ so that, for $s = 1,\ldots,N$,

$$\frac{\partial E(\bar{\alpha},\bar{\beta})}{\partial \alpha_{ss}} = \int_{A_{s-1}}^{A_s} \pi_1(x - A_{s-1}) \cdot (x - A_{s-1})f(x)dx + \int_{A_s}^{A_N} \pi_1(\bar{a}_s) \cdot \bar{a}_s f(x)dx = \mu_s, \text{ since } \alpha_{ss} > 0;$$

$$\frac{\partial E(\bar{\alpha},\bar{\beta})}{\partial \alpha_{sr}} = \int_{A_{s-1}}^{A_s} \pi_1(0)(x - A_{s-1})f(x)dx + \sum_{t=s+1}^{r-1} \int_{A_{t-1}}^{A_t} \pi_1(0)\bar{a}_s f(x)dx + \int_{A_{r-1}}^{A_r} \pi_1(x - A_{r-1})\bar{a}_s f(x)dx + \int_{A_r}^{A_N} \pi_1(\bar{a}_r)\bar{a}_s f(x)dx \leq \mu_r$$

$$\text{for } r > s$$

But $r > s$, $E(\bar{\alpha},\bar{\beta}) = E^A$ implies $\bar{a}_r < \bar{a}_s$ by Proposition 2, and $\pi_1(a_i)$ decreasing in a_i implies on a term-by-term basis that

$$\frac{\partial E(\alpha,\beta)}{\partial \alpha_{sr}} > \frac{\partial E(\alpha,\beta)}{\partial \alpha_{ss}}$$

This contradicts the first-order conditions, hence the appropriative allocation is nonoptimal.

APPENDIX C

Given competitive markets in water rights and diversion facilities, each firm solves the problem

$$\max E^j\pi = \sum_{i=1}^{N} \int_{A_{i-1}}^{A_i} \pi[\alpha_{ij}(x - A_{i-1}) + \sum_{k=1}^{i-1} \alpha_{kj}\bar{a}_k, \bar{a}_j] f(x)dx + [1 - F(A_N)]\pi(\sum_{k=1}^{N} \beta_{kj}\bar{a}_k,\bar{a}_j) + p_j - \sum_{i=1}^{N} p_i\alpha_{ij} + q_j - \sum_{i=1}^{N} q_i\beta_{ij}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \alpha_{ij}a_i \leqq \sum_{i=1}^{N} \beta_{ij}\bar{a}_i$$

with first-order conditions

$$\frac{\partial E^j\pi}{\partial \alpha_{ij}} = \int_{A_{i-1}}^{A_i} \pi_1[\alpha_{ij}(x - A_{i-1}) + \sum_{k=1}^{i-1} \alpha_{kj}\bar{a}_k](x - A_{i-1})f(x)dx + \sum_{r=i+1}^{N} \int_{A_{r-1}}^{A_r} \pi_1[\alpha_{rj}(x - A_{r-1}) + \sum_{k=1}^{r-1} \alpha_{kj}\bar{a}_k]\bar{a}_i f(x)dx - p_i - \lambda_j\bar{a}_i = 0$$

$$\text{for } \alpha_{ij} > 0, i = 1,\ldots,N$$

$$\frac{\partial E^j}{\partial \beta_{ij}} = [1 - F(A_N)] \pi_1(\sum_{k=1}^{N} \beta_{kj}\bar{a}_k)\bar{a}_i - q_i + \lambda_j\bar{a}_i = 0$$

for $\beta_{ij} > 0$, $i = 1,\ldots,N$, where λ_j is the multiplier associated with the constraint $\Sigma_{i=1}^{N} \alpha_{ij}\bar{a}_i \leq \Sigma_{i=1}^{N} \beta_{ij}\bar{a}_i$. Note that the allocation $\alpha_{ij} = \beta_{ij} = 1/N$ $i, j = 1,\ldots, N$ which maximizes $\Sigma_{j=1}^{N} E^j\pi$ also satisfies the first-order conditions for any j, given that p_i and q_i satisfy equations (1) and (2) of the main text.

REFERENCES

H. S. Burness and J. P. Quirk, "The Colorado River: Water Rights Institutions, Priorities, and Allocations," Environ. Quality

Lab. manuscript, California Instit. Technology, Dec. 1976.

——— **and** ———, "Water Rights and Optimal Reservoir Management," soc. sci. work. paper no. 165, California Instit. Technology, Dec. 1977.

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

Richard L. Dewsnut and Dallice W. Jensen, *Summary Digest of State Water Laws*, Washington 1973.

J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.*, Mar. 1969, *59*, 25–34.

Charles J. Meyers, "The Colorado River," *Stanford Law Rev.*, Nov. 1966, *19*, 1–75.

——— **and A. Dan Tarlock,** *Water Resource Management*, Mineola 1971.

J. Milliman, "Water Law and Private Decision Making: A Critique," *J. Law Econ.*, Oct. 1959, *2*, 41–63.

J. P. Quirk and R. Saposnik, "Admissibility and Measureable Utility Functions," *Rev. Econ. Stud.*, Feb. 1962, *29*, 140–46.

Optimal Pricing with Intermodal Competition

By RONALD R. BRAEUTIGAM*

The regulation of multiproduct enterprises has created some difficult problems for regulators, particularly where common costs of production are present and where entry may be allowed in one or more of the markets served by such a firm. This paper concentrates on aspects of economic efficiency in pricing with multiproduct firms and intermodal competition. It extends the work of William Baumol and David Bradford on efficient pricing with multiproduct monopoly to the case where intermodal competition is present. A set of rules is developed, showing how second best prices deviate from marginal cost when economies of scale are present. The paper shows why these rules may be difficult to implement in some cases, with a direct application to the case of surface freight transport, and then suggests a variation in the theory of second best which may be useful given those difficulties.

I. Fair Prices and Ramsey Optimality

At least three factors contribute to the difficulty of solving the pricing problem in regulated firms. First, it may be the case that when price is set equal to marginal cost in each of the markets served by a firm, profits would be negative. Hence, some deviation of price from marginal cost is required if the firm is to break even. Second, there may be costs of production which are shared by two or more services in the production process, so that it is impossible to assign costs to services in an unambiguous manner. Finally, there may be other firms participating in some of the markets served by a multiproduct firm. In such a case pricing policies may affect market structure, and the two should not be treated independently.

Through the processes of administrative law, regulatory decisions have emphasized "fairness" in pricing rather than economic efficiency. There are many possible concepts which could be used to define fair prices (see, for example, Gerald Faulhaber, 1972; Edward Zajac). In judging whether prices are fair, regulators have historically tended to allocate shared costs first, and then to require that the price charged for any service generate revenues which cover the portion of shared costs allocated to that service plus all costs that can unambiguously be attributed to that service. The rather lengthy proceedings of the Federal Communications Commission (*FCC*) and of the Interstate Commerce Commission (*ICC*) address the manner in which shared costs are to be allocated. Importantly, the prices set as a result of this process may bear no direct relationship to economic efficiency.

The issues of efficient pricing for a multiproduct firm have been examined in the classic article by Baumol and Bradford. Briefly, they developed rules for second best pricing in a firm which would earn negative profits, if price were equal to marginal cost in each market. The pricing rules maximize economic efficiency (as measured by the sum of producer's and consumer's surplus) subject to a constraint which allows the firm to break even.[1]

Suppose a firm produces n commodities in quantities $x_1, \ldots, x_n$, and, for simplicity,

*Assistant professor of economics and the Transportation Center, Northwestern University. This paper was partially sponsored by the National Science Foundation. I would like to thank Al Klevorick, John Ledyard, John Panzar, John Roberts, and Robert Willig for their helpful suggestions in review of an earlier version; George Borts and John Chilton have provided especially helpful comments in the final stages.

[1] In the paper of Baumol and Bradford, the authors asserted that the form of the utility function being maximized was unspecified However, Herbert Mohring demonstrated that the unspecified utility function actually had the properties of the consumer's surplus measure. For more on this measure, see Robert Willig.

that the demands for the commodities are independent of one another. Assume also that the cost function for the production process can be represented by $C(x_1, x_2, \ldots, x_n)$. Then the second best prices $(p^1, p^2, \ldots, p^n)$ are those which satisfy equations (1) and (2).

$$(1)\qquad R^i \triangleq \left[\frac{p^i - (\partial C/\partial x_i)}{p^i}\right]\epsilon_p^i = \left[\frac{p^j - (\partial C/\partial x_j)}{p^j}\right]\epsilon_p^j \triangleq R^j, \qquad \forall ij$$

$$(2)\qquad \sum_{i=1}^{n} p^i x_i - C = 0$$

where ϵ_p^i represents the price elasticity of demand in the ith market.

Equation (2) represents a condition in which the firm is breaking even (total revenues equal total cost), and equation (1) represents the well-known rule that in each market the amount by which price deviates from marginal cost is inversely related to the price elasticity of demand. The numbers R^i and R^j are sometimes called Ramsey numbers, based on the work of Frank Ramsey which was further developed by Baumol and Bradford. The theory has been extended to cover the case in which the demands are interdependent, resulting in a slightly more complicated form for the Ramsey numbers.[2] The basic idea remains unchanged in characterizing second best, namely, the Ramsey numbers are equal in all markets and the firm is earning zero profits.

It is important to note that prices which satisfy equations (1) and (2) may not generally satisfy all of the possible definitions of "fair" prices examined by Zajac and Faulhaber.[3] There is an essential difference between the approaches to pricing taken by regulators and by Baumol and Bradford. Regulators tend to allocate shared costs first, and then judge prices based on that allocation as described above. In the work of Baumol and Bradford, efficient prices are based on marginal costs and conditions of demand. No prior allocation of shared costs is required. (It is possible to determine how shared costs should be allocated in order to reach second best once the efficient prices have been found, but the allocation is done *ex post* instead of *ex ante*.)

As developed by Baumol and Bradford, the theory of second best applies only to a firm which has a monopoly in each of its markets. One of the important gaps which has not yet been filled is the case of intermodal competition. Several questions may be posed. Is the notion of Ramsey optimality useful with intermodal competition? If so, what do the Ramsey numbers look like? What particular kinds of difficulties might be expected in an application of the theory, and what modifications in the notion of second best may be of interest as a result of this line of investigation? These questions are addressed in the next section, using the regulation of surface freight transportation as an exemplary context.

II. Second Best and Intermodal Competition

The regulation of surface freight transportation has posed perplexing problems for the *ICC*, in large part because of the growth of competition among the various modes of transport, especially since about 1930. Among other things, not all of the modes are characterized by economies of scale. According to Dudley Pegrum, "railroads and pipelines have the basic economic characteristics of public utilities and are what economists call natural monopolies; motor, water, and air transport exhibit the features of competitive industries" (p. 25).

Freight transportation in this country has certain distinctive features which lead us to concentrate on the interactions among rail, motor, and water carriers. First, air carriers primarily provide passenger service. In 1976 passenger service accounted for 86 percent of the revenue for domestic air trunk lines. In the same year air cargo (which includes freight, express, and mail traffic) accounted for only 0.2 percent of all intercity private and for-hire ton miles

[2]One place (and there are quite probably others) in which modified Ramsey numbers are derived for a multiproduct firm with increasing returns to scale and interdependent demands is in the author (1976).

[3]See the Zajac paper for some clear examples of this point.

carried in this country.[4] Second, pipelines "constitute a highly specialized form of transportation for the movement of products in liquid or gaseous form" (see Pegrum, p. 43). Because of this special nature of the service they provide and their apparent economies of scale, the regulation of oil and gas pipelines could be treated separately, under the jurisdiction of either the *ICC* or the Department of Energy.[5]

The remaining three modes employ greatly differing technologies to provide services which can be viewed as imperfect substitutes for one another. It should be noted that the issue of economies of scale in railroads is not a closed matter. Several empirical studies have been made to test for the existence of economies of scale, with results that have generally been mixed. For example, Lawrence Klein used 1936 data to find statistically significant, though modest, economies of scale. However, studies by George Borts and Zvi Griliches have concluded that even if scale economies are present for smaller railroads, they are not prevalent in the larger ones.

It is not the purpose of this paper to critique these empirical studies. Rather, the intent is to examine how second best prices might be set if one of the modes has scale economies (and railroads appear to be the most likely candidate) and the other modes (water and motor) do not.[6] If none of these modes has increasing returns to scale, the basis for any regulation at all should be examined. If any of the modes do have scale economies, then the questions addressed in this paper are appropriate ones to examine.

[4]See Donald Harper, pp. 300, 301. The exclusion of air freight is made here primarily for simplicity. Many of the arguments developed later on could be extended to encompass air freight simply by letting this mode be included as one of the m modes in the model to be developed.

[5]Thomas Moore suggests the separate regulation of oil pipelines, and recognizes the natural monopoly characteristics of this mode.

[6]For a recent empirical examination of returns to scale in the regulated motor carrier industry, see Ann Friedlaender. She concludes that in the absence of entry and operating restrictions it is likely that the trucking industry would be competitively organized, and that the efficiently sized firm would be quite small relative to the relevant market.

Let us construct a model of intermodal competition using the following assumptions:

1) There are m modes which provide transport services between two points. Only one of these modes (mode 1) is characterized by economies of scale. In other words, if the services provided by mode 1 were all priced at marginal cost, the profits for the firm would be negative.

2) There are many suppliers of transport service in each of the other modes, so that each of the modes $2, \ldots, m$ is essentially competitive. It is assumed that with free entry the supply of transport services in each of these modes would be perfectly elastic.

3) Each mode may transport any or all of n commodities. Let

i = a modal index, $i = 1, \ldots, m$

j = a commodity index, $j = 1, \ldots, n$

x_{ij} = the amount of commodity j transported by mode i

4) All carriers of mode i provide identical service in the transport of commodity j. Restated, this means that there is intramodal service homogeneity in the carriage of a particular commodity.

5) It is assumed that the demand schedules can be represented as inverse demand functions, and that the demand for transportation of commodity j via any mode is independent of the demand for transportation of commodity $k (k \neq j)$ via any mode. Formally, let

$$p^{ij} = p^{ij}(x_{1j}, x_{2j}, \ldots, x_{mj}), \quad i = 1, \ldots, m; \; j = 1, \ldots, n$$

where p^{ij} represents the (inverse) demand for the transport of commodity j via mode i.

6) There is intermodal service differentiation. In transporting commodity j, carriers of one mode will provide service which differs from the service of carriers of other modes. This recognizes that motor carriers, water carriers, and railroads may differ in the speed of transport, reliability, and in other aspects of service quality. Assume that the services provided by the different modes can be characterized as weak

gross substitutes with the property $\partial p^{ij}/\partial x_{kj} < 0, i \neq k$.

In addition, let S^{ij} = the price corresponding to the (perfectly elastic) supply schedule for mode i in the provision of service j, and

$$C^1 = C^1(x_{11}, x_{12}, \ldots, x_{1n}; \text{factor prices})$$

be the total cost function for mode 1. Factor prices are assumed constant, so reference to them is suppressed throughout the rest of this paper.

Finally, assume that there are zero income effects associated with the demand functions p^{ij}, so that a measure of the gross benefits from the provision of $(x_{11}, \ldots, x_{1n}; x_{21}, \ldots, x_{2n}; \ldots; x_{m1}, \ldots, x_{mn})$ is defined by G, where,

$$(3)\quad G = \sum_{j=1}^{n} \{\int_{w=0}^{x_{1j}} p^{1j}(w, 0, \ldots, 0)dw + \int_{w=0}^{x_{2j}} p^{2j}(x_{1j}, w, 0, \ldots, 0)dw + \ldots + \int_{w=0}^{x_{mj}} p^{mj}(x_{1j}, \ldots, x_{m-1,j}, w)dw\}$$

If G were maximized, given a set of prices p^{ij}, then x_{ij} would be chosen so that $\partial G/\partial x_{ij} = p^{ij}$.

We can now write a function T, which measures the sum of consumer's and producer's surplus associated with any level of service:

$$(4)\quad T = G - C^1 - \sum_{i=2}^{m} \sum_{j=1}^{n} S^{ij} x_{ij}$$

We are now ready to examine the nature of a second best operating point when the regulator is able to select the levels of x_{ij} for all i and j. Note that the question of second best is of interest, since if the regulator attempted to reach first best, we would have

$$(5)\quad \frac{\partial T}{\partial x_{1j}} = p^{1j} - \frac{\partial C^1}{\partial x_{1j}} = 0, \forall j$$

which means that the mode with economies of scale would be earning negative profits.

If the regulator wants to set the levels of x_{ij} to maximize efficiency while allowing the mode 1 firm to break even, then formally it would find a solution to the following problem, which I call totally regulated second best (*TRSB*):

$$(6)\quad \max_{x_{ij}, \forall i,j} T = G - C^1 - \sum_{i=2}^{m} \sum_{j=1}^{n} S^{ij} x_{ij}$$

$$\text{subject to} \quad \sum_{j=1}^{n} p^{1j} x_{1j} - C^1 \geq 0$$

Define L as follows, where λ is the nonnegative Lagrangean associated with the break-even constraint:

$$(7)\quad L = G - C^1 - \sum_{i=2}^{m} \sum_{j=1}^{n} S^{ij} x_{ij} + \lambda\left(\sum_{j=1}^{n} p^{1j} x_{1j} - C^1\right)$$

Among the first-order conditions are

$$(8)\quad \frac{\partial L}{\partial x_{1j}} = p^{1j} - \frac{\partial C^1}{\partial x_{1j}} + \lambda\left(\frac{\partial p^{1j}}{\partial x_{1j}} x_{1j} + p^{1j} - \frac{\partial C^1}{\partial x_{1j}}\right) \leq 0,$$

$$x_{1j} \geq 0, x_{1j} \frac{\partial L}{\partial x_{1j}} = 0; \qquad j = 1, \ldots, n$$

and

$$(9)\quad \frac{\partial L}{\partial x_{ij}} = p^{ij} - S^{ij} + \lambda\left(\frac{\partial p^{1j}}{\partial x_{ij}} x_{1j}\right) \leq 0,$$

$$x_{ij} \geq 0, x_{ij} \frac{\partial L}{\partial x_{ij}} = 0;$$

$$i = 2, \ldots, m; j = 1, \ldots, n$$

Equation (8) can be rewritten (assuming $x_{1j} > 0$)

$$(10)\quad \left[\frac{p^{1j} - (\partial C^1/\partial x_{1j})}{p^{1j}}\right]\left[\frac{p^{1j}}{(\partial p^{1j}/\partial x_{1j})x_{1j}}\right] = -\frac{\lambda}{1+\lambda}; \qquad j = 1, \ldots, n$$

where the second term on the left-hand side is the reciprocal of the quantity elasticity of demand for x_{1j}. One could think of the expression on the left-hand side of equation (10) as a modified Ramsey number which will be equal for all values of j, since $-\lambda/(1 + \lambda)$ does not vary with j.

However, equation (9) must also be satisfied, and here we encounter a potential administrative nightmare. Note that $\partial L/\partial x_{ij} = 0$ whenever $x_{ij} > 0$. Since the demands for x_{1j} and $x_{ij} (i \geq 2)$ are not independent, the term

$$\lambda \frac{\partial p^{1j}}{\partial x_{ij}} x_{1j}$$

is not zero. In fact, as long as x_{1j} and x_{ij} are weak gross substitutes for one another, this term will have a negative sign. Hence, a second best solution in which x_{ij} is positive would occur only when the price exceeds marginal cost in modes 2, ..., m. There are two effects working against each other which make this property interesting. Heuristically, there is some loss in efficiency which occurs in the markets served by modes 2, ..., m because price is greater than marginal cost. However, the higher prices in modes 2, ..., m lead to increased demands (and more consumer's surplus) for the services provided by mode 1. Equation (9) implies that the second effect exceeds the first.

In principle one could calculate Ramsey numbers for modes 2, ..., m which would equal the number $-\lambda/(1 + \lambda)$ from equation (10), although the form of these numbers is more complex than the modified Ramsey number in that equation. Using the assumption of zero income effects, by Hotelling's integrability condition we have

$$(11) \qquad \frac{\partial p^{1j}}{\partial x_{ij}} = \frac{\partial p^{ij}}{\partial x_{1j}}, \qquad \forall i,j$$

From equations (9) and (11) it follows that

$$(12) \qquad \frac{(p^{ij} - S^{ij})/p^{ij}}{(\partial p^{ij}/\partial x_{1j})(x_{1j}/p^{ij}) - (p^{ij} - S^{ij})/p^{ij}} = -\frac{\lambda}{1 + \lambda}; i = 2, \ldots, m; j = 1, \ldots, n$$

The numerator of the left-hand side represents the amount by which the price of x_{ij} would exceed marginal cost, stated as a fraction of the price itself. A similar expression appears in the second term of the denominator. The first term of the denominator represents the cross elasticity of the inverse demand p^{ij} with respect to the quantity x_{1j}.

The achievement of the second best would then require that

1) Mode 1 earns zero economic profit.

2) Prices are set so that the modified Ramsey numbers for all modes in all markets are equal to one another. The modified Ramsey numbers for mode 1 are defined by equation (10), and for all other modes are as shown in equation (12).

3) Since price exceeds marginal cost in the markets served by modes 2, ..., m, the regulator would have to prevent free entry in those markets, or else impose a set of taxes designed to hold tariffs above marginal costs.

There can be little doubt that the regulatory scheme just outlined represents an enormous regulatory undertaking. Some might argue that there is a striking similarity between the outlined program and the actual kind of regulation we observe in freight transportation presently. After all, regulators do adjudge the reasonableness of prices (tariffs) for all regulated modes, and in addition control conditions of entry through certificates required of common carriers wishing to provide service over particular routes. One could even argue that through a consideration of "value of service" in pricing, regulators attempt to require higher tariffs on commodities with more inelastic demands, and that this is generally consistent with the guidelines suggested by rules such as those of equations (10) and (12).

However, one would be hard pressed to carry the analogy much further. It would be an understatement to say that the data requirements for the outlined program are great. In fact, the information required on the numerous cross elasticities of demand alone is enough to make the outlined program quite unwieldy.

Unfortunately, even if we were to commit ourselves to the quest for second best, we are likely to encounter other difficulties at least as important as the information requirements. The case of freight transportation serves well to illustrate this point. Suppose that mode 2 represents regulated motor carriage, and that a regulator desires to hold tariffs above marginal costs for this mode. Under their present statutory powers, regulators are not empowered to impose taxes, even if they had sufficient information to determine the levels of the taxes required.

If a program of taxes cannot be implemented, then a regulator may attempt to limit entry in order to hold tariffs above marginal costs. However, the presence of an unregulated portion of the motor carrier industry, as we have in this country, may present an overwhelming problem. For example, if prices are held above marginal costs for regulated carriers, shippers who would otherwise have used regulated motor carriers will have incentives to buy their own trucks for the purposes of hauling their own commodities. Such private haulage is not regulated, and thus could not be prevented by the regulations applying to common carriers. As a result, although the intent of regulation is to proscribe entry, the probable effect would simply be to change the form of entry to circumvent the regulation.

These difficulties lead to a search for some modified form of second best solution that requires less information on the part of regulators. We want to avoid the entry control problems described above for modes which do not experience significant technological barriers to entry, such as economies of scale. One rather interesting candidate for examination would be a regulatory program that allows the modes without economies of scale ($i = 2, \ldots, m$) to clear their respective markets, and concentrates on the prices set by the mode with economies of scale. In terms of administration, regulators would not have to set the $n(m - 1)$ tariffs (or quantities) for the modes without increasing returns to scale, and in addition would not concern themselves with the thorny problem of entry control in those modes. The administration of regulation under this scenario would be much simplified.

There are other reasons why such a program, which I will call partially regulated second best, might be of interest. In recent years there has been much debate over the extent to which regulation is actually needed in freight transport. Since 1930 technological changes have made the transport of freight by motor and water carriers economically viable on a large scale. To control the interactions among modes, the hand of regulation has been extended repeatedly. However, it is sometimes argued that the presence of viable alternative modes may mean that with less regulation market forces might work quite well in making many of the resource allocation decisions now made by regulators. The information requirements and welfare properties of such a program will be addressed in the next section.

III. Partially Regulated Second Best (*PRSB*)

There are two ways one could formalize the suggested concept of partially regulated second best. First, a set of market-clearing constraints for modes $2, \ldots, m$ could be appended to the problem defined in equation (6). Then the optimal constraints would be

$$(13) \quad p^{ij} - S^{ij} = 0; \qquad i = 2, \ldots, m; \; j = 1, \ldots, n$$

There would be a Lagrange multiplier associated with the break-even constraint (λ), and one associated with each market-clearing condition (μ^{ij}). Unfortunately, this approach does not easily lend itself to the derivation of a set of expressions equal across markets that avoids explicit reference to the values of the Lagrange multipliers λ and μ^{ij}. Thus an important advantage of the approach used by Baumol and Bradford is lost.

Fortunately, there is a second method which will lead us to a surprisingly simple result. The set of expressions of equation (13) can be used to yield representations of $x_{ij}(i \geq 2)$ as implicit functions of x_{1j}. We can then characterize the partially regulated second best problem as the same system as equation (6), except that the set of decision variables includes only the x_{1j} terms, and not x_{ij} $(i \geq 2)$.

A derivation of the pricing rules for the partially regulated second best problem appears in the Appendix. At an interior optimum (in which x_{1j} and x_{1k} are positive), the following conditions must be satisfied:

$$(14) \qquad \sum_{j=1}^{n} p^{1j} x_{1j} - C^1 = 0$$

$$(15)\quad \left[\frac{p^{1j} - (\partial C^1/\partial x_{1j})}{p^{1j}}\right] \epsilon_p^{1j} = \left[\frac{p^{1k} - (\partial C^1/\partial x_{1k})}{p^{1k}}\right] \epsilon_p^{1k}$$

$$\text{for } j = 1,\ldots,n;\ k = 1,\ldots,n$$

The partially regulated second best prices for the case *with intermodal competition* are set according to the same rules as the ones developed by Baumol and Bradford for a multiproduct monopoly. The Ramsey numbers defined in equation (15) depend only on local information on price, marginal cost, and the price elasticity of demand for the first mode.[7]

Upon reflection, these results do have an intuitive appeal. The pricing rules of Baumol and Bradford are conceptually appropriate when goods or services produced by mode 1 have demands which are independent of the demands for goods or services produced by other firms. In other words, the multiproduct firm must monopolize all of its markets. However, suppose there are other products whose demands interact with the outputs of mode 1. Then one could describe second best for the whole set of these products, as I have done earlier. The results say that if one mode has economies of scale, it may be efficient (second best) to alter the market-clearing outcomes for other modes, even if those modes serve markets which are potentially quite competitive.

There are several reasons why a regulator may not even attempt to specify a program of total regulation leading to second best. Regulators may perceive the interactions among the demands for products of mode 1 and other modes to be small, or they may simply be unaware of the interaction. They may also recognize the potentially very large information and administrative requirements for such a program, or the difficulties in controlling entry as effectively as would be required. There may be other reasons for which regulators may explicitly decide to let the markets clear for those modes which are essentially competitive.[8] In any one of these cases partially regulated second best becomes an interesting candidate for efficient pricing.

Suppose that a regulator mistakenly thinks that the multiproduct firm has a monopoly in its markets. Then it might determine second best prices by the rules of Baumol and Bradford. However, this is equivalent to a situation in which the regulator explicitly recognizes the interdependence of mode 1 products with the outputs of other (competitive) modes, but decides to allow the markets for other modes to clear. Thus, the connection between partially regulated second best and the rules of Baumol and Bradford is drawn more clearly.

IV. A Comparison of Welfare Properties

To illustrate the basic properties of several interesting operating points with intermodal competition, it is useful to consider a special case for which a graphical exposition is possible. Assume that there are only two modes, mode 1 with economies of scale (as before), and mode 2 which lacks scale economies. Only one basic kind of service is provided by each mode. The service provided by mode 1 is differentiated from the service of mode 2. However, all firms in mode 2 provide a homogeneous service. We have retained the assumption of intermodal service differentiation and intramodal service homogeneity. For the purpose of the illustration, mode 2 will be characterized as having a supply schedule which may or may not be perfectly elastic. In other words, the supply schedule can be written as $S^2(x_2)$, where S^2 represents the price at which x_2 units of service would be provided by producers in mode 2. As long as the mode 2 market clears, we have

[7]The more complicated case in which all commodities transported by all modes have interdependent demands could be approached in the same way as for the simpler case developed in this paper, in a manner similar to the extension of the Baumol-Bradford framework to the case of interdependence as described earlier.

[8]A regulator may find it desirable to impose other constraints on the system. In principle, these constraints could be appended to a model which has as its objective function the maximization of surplus to find efficient prices given these additional constraints.

(16) $$p^2(x_1, x_2) - S^2(x_2) = 0$$

which implies that

(17) $$\frac{dx_2}{dx_1} = -\frac{(\partial p^2/\partial x_1)}{(\partial p^2)/\partial x_2 - (\partial S^2/\partial x_2)} < 0$$

The property that $dx_2/dx_1 < 0$ holds when x_1 and x_2 are weak gross substitutes and when the demand for x_2 is more negatively sloped than the supply curve. Since x_1 and x_2 are assumed to be weak gross substitutes and since mode 2 is assumed not to have increasing returns to scale, the inequality in equation (17) is implied. A locus of points satisfying equation (16) is represented in Figure 1 by the curve AE. Point A corresponds to the point at which only mode 2 serves the market. The negative slope of AE follows from equation (17).

We may also represent the sum of consumer's and producer's surplus by T,

(18) $$T = \int_{w=0}^{x_1} p^1(w,0)dw + \int_{w=0}^{x_2} p^2(x_1,w)dw - C^1(x_1) - \int_{w=0}^{x_2} S^2(w)dw$$

As a result, iso-surplus curves will have the slope

(19) $$\frac{dx_2}{dx_1} = -\frac{p^1 - (\partial C^1/\partial x_1)}{p^2 - S^2}$$

Note that along the curve AE the iso-surplus curves are vertical, since $p^2 - S^2 = 0$. Also, T increases along AE as x_1 increases up to a level of output at which p^1 equals the marginal cost of producing x_1, where T reaches its maximum. Thus, along AE

(20) $$dT = \left(p^1 - \frac{\partial C^1}{\partial x_1}\right)dx_1 + (p^2 - S^2)dx_2 = \left(p^1 - \frac{\partial C^1}{\partial x_1}\right)dx_1$$

Let point E represent the level of output at which price equals marginal cost for mode 1 along AE. Since T is maximized here, E represents a first best operating point. The iso-surplus curves around E shown in Figure 1 will have values such that $T_E > T_D > T_C > T_B$.

The profit of mode 1 can be expressed as Π^1, where

(21) $$\Pi^1 = p^1(x_1,x_2)x_1 - C^1(x_1)$$

The iso-profit curves for mode 1 will have the slope

$$\frac{dx_2}{dx_1} = -\frac{p^1 + (\partial p^1/\partial x_1)x_1 - (\partial C^1/\partial x_1)}{(\partial p^1/\partial x_2)x_1}$$

Since x_1 and x_2 are weak gross substitutes, the sign of the slope will be positive when the marginal revenue for x_1 exceeds the marginal cost of x_1 (for levels of output less than the profit-maximizing level, given x_2), and negative when the converse is true. The shapes of these iso-profit curves are shown in Figure 1. The ordering of the profit levels can be seen by noting that given any level of x_1, the profit of mode 1 will increase when x_2 decreases, i.e.,

(22) $$\frac{\partial \Pi^1}{\partial x_2} = \frac{\partial p^1}{\partial x_2}x_1 < 0$$

Let us now put all of this together. Figure 1 is drawn to reflect the case in which it is possible for mode 1 to at least break even for some operating points when the market for x_2 clears. Suppose both modes were unregulated, and that mode 1 chooses the highest iso-profit curve it can attain given that mode 2 will clear. Then this point of no regulation is shown at B.

If a regulator wants to maximize efficiency while allowing mode 2 to clear and mode 1 to just break even, it would choose point C. Point C thus represents a $PRSB$

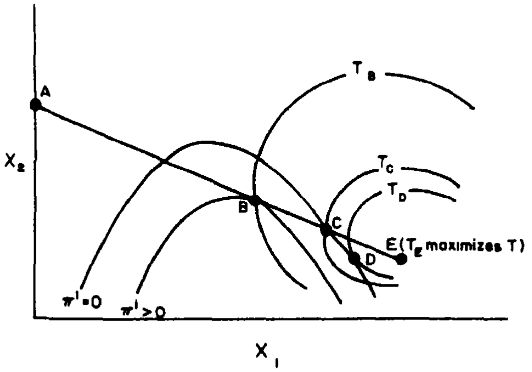


FIGURE 1

point as determined by equations (14) and (15). If a regulator chooses to maximize efficiency while allowing mode 1 to break even, and is willing to undertake the control of quantities (or tariffs) and entry in mode 2, then it would strive to reach the *TRSB* point, *D*. At *D*, the iso-surplus curve T_D is tangent to the zero iso-profit curve for mode 1. Since the slope of the iso-profit curve for mode 1 is not vertical at that point, *D* must be located below the curve *AE*. This points out that at a totally regulated second best solution, the market for mode 2 will not be clearing, and either a set of taxes or direct entry control will be necessary.

The relationships between the iso-profit curves for mode 1 and the market-clearing locus *AE* could be other than as depicted in Figure 1. For example, if mode 1 cannot break even at any market-clearing price in mode 2, then Figure 2 is appropriate. There exists no *PRSB* point (point *C* in Figure 1) and no totally unregulated point where mode 1 is profitable (such as point *B* in Figure 1). Mode 1 can only break even (in the absence of a subsidy) when mode 2 is prevented from clearing its market, and the most efficient point of operation where mode 1 breaks even is the *TRSB* point *D*.

In between the situations shown in Figures 1 and 2 is the one in which an unregulated mode 1 would just barely be able to break even, such as in Figure 3. If mode 1 could just earn zero profit in this case, then the *PRSB* point *C* and totally unregulated point *B* would coincide. In this case the

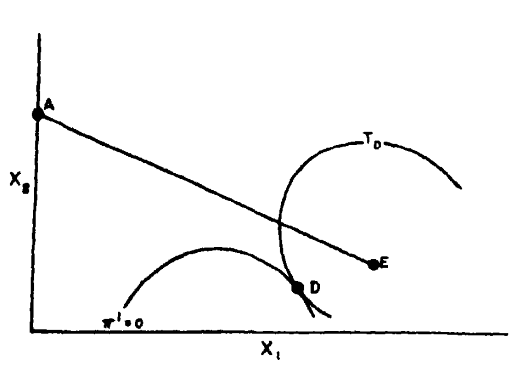


FIGURE 2

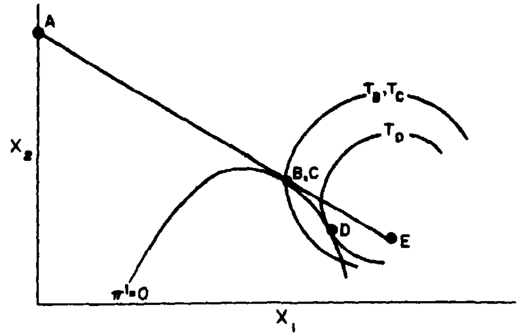


FIGURE 3

Ramsey numbers of equation (15) would be equal to minus one.[9] This suggests that if it is expected that without any regulation only small economic profits would be earned by the mode with economies of scale, then an unregulated system would achieve nearly the same efficiency as *PRSB*, and without incurring the administrative costs of the latter.

V. Conclusions

This paper has shown how the theory of second best can be extended from the work of Baumol and Bradford to a case in which intermodal competition exists. I have derived rules characterizing second best under a form of intermodal competition which may resemble what we observe in freight transportation in this country. There are at least two major problems which regulators should anticipate if they attempt to reach second best when all modes are regulated. First, there appears to be a large amount of information required to use the rules which are derived. Some of this information may be difficult to obtain, particularly since cross elasticities of demand are important. Second, the achievement of second best may involve a departure of prices from marginal cost even for modes which would be essentially competitive in the absence of regulation. The realization of those prices would require either the imposition of taxes or the

[9] For a further development of this point, see Appendix B.

careful control of entry in markets that may not easily lend themselves to such control, such as with motor carrier freight transportation.

These potential difficulties led to the investigation of a modified form of second best. This form, called partially regulated second best, does not require the direct regulation of modes which appear to be essentially competitive. Prices are specified for a mode with economies of scale, and these prices are designed to maximize efficiency subject to conditions which allow that mode to break even, while the modes without increasing returns to scale are clearing their markets. The rules derived for *PRSB* have the same form as the ones developed by Baumol and Bradford for the case without intermodal competition. If total regulation were costless and effective in achieving second best, then *TRSB* would be more efficient than *PRSB*. However, if the former is costly to achieve (because of large information requirements) or is otherwise difficult to reach (for example, because of the inability to effectively control entry), then *PRSB* may become an attractive alternative.[10]

I have compared both of these alternatives to a third one in which there is no regulation of tariffs or entry on any of the modes. Again, if regulation were costless and effective, both forms of second best will achieve greater economic efficiency than the unregulated system, as long as the mode with economies of scale could earn positive economic profits if it were not regulated. However, if the level of positive profits attainable without regulation is near zero, then the efficiency achieved without regulation may be quite close to that reached under *PRSB*. Once again, the information-gathering and administrative costs associated with *PRSB* may be large relative to the case with no regulation. The qualitative nature of the tradeoff between administrative costs and attainable efficiency is clear; however a quantitative determination depends on characteristics specific to an industry. For the case of freight transport, the quantitative determination remains for further work.

[10] I have examined the rate structure of the *U.S.* railroad industry in 1961 to see how the structure differed from *PRSB* rates. The Ramsey numbers corresponding to equation (15) in the text were determined to be as follows: −0.075 for products of agriculture; −0.06 for animals and products; −0.072 for products of mines; −0.135 for products of forests; −0.224 for manufactured and miscellaneous commodities. These results suggest that agricultural tariffs may have been too low relative to tariffs for manufacturing and miscellaneous products to be economically efficient. For a development and discussion of this, see my forthcoming paper.

Appendix A—Derivation of *PRSB* Pricing Rules (Equations (14) and (15))

The *PRSB* problem can be written

$$\max_{(x_{11},\ldots,x_{1n})} T = G - C^1 - \sum_{i=2}^{m}\sum_{j=1}^{n} S^{ij}x_{ij}$$

subject to $$\sum_{j=1}^{n} p^{1j}x_{1j} - C^1 \geq 0$$

where λ will be the Lagrange multiplier associated with the break-even constraint. Under this formulation the x_{ij} variables are implicit functions of the x_{1j} variables, i.e., $x_{ij} = x_{ij}(x_{1j})$, $2 \leq i \leq m$, $1 \leq j \leq n$. When the constraint is binding, equation (14) holds.

At an interior optimum ($x_{1j} > 0$), the first-order conditions require that

$$\text{(A1)}\quad (1+\lambda)\left(p^{1j} - \frac{\partial C^1}{\partial x_{1j}}\right) = -\lambda x_{1j}\left[\frac{\partial p^{1j}}{\partial x_{1j}} + \sum_{i=2}^{m}\frac{\partial p^{1j}}{\partial x_{ij}}\frac{dx_{ij}}{dx_{1j}}\right]$$

The key to an easy interpretation of this condition lies in the meaning of the term in brackets on the right-hand side. We define two matrices, $[B^j]$ and $[A^j]$ follows:

$$[B^j] \triangleq \begin{bmatrix} \partial p^{2j}/\partial x_{2j} & \cdots & \partial p^{2j}/\partial x_{mj} \\ \partial p^{3j}/\partial x_{2j} & \cdots & \partial p^{3j}/\partial x_{mj} \\ \vdots & \vdots & \vdots \\ \partial p^{mj}/\partial x_{2j} & \cdots & \partial p^{mj}/\partial x_{mj} \end{bmatrix}$$

$$[A^j] \triangleq \begin{bmatrix} \frac{\partial p^{1j}}{\partial x_{1j}} & \frac{\partial p^{1j}}{\partial x_{2j}} & \cdots & \frac{\partial p^{1j}}{\partial x_{mj}} \\ \frac{\partial p^{2j}}{\partial x_{1j}} & \frac{\partial p^{2j}}{\partial x_{2j}} & \cdots & \frac{\partial p^{2j}}{\partial x_{mj}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial p^{mj}}{\partial x_{1j}} & \frac{\partial p^{mj}}{\partial x_{2j}} & & \frac{\partial p^{mj}}{\partial x_{mj}} \end{bmatrix} = $$

$$\triangleq \begin{bmatrix} \frac{\partial p^{1j}}{\partial x_{1j}} & a^j_{1.} \\ a^j_{.1} & B^j \end{bmatrix}$$

Equation (A2) follows from a straightforward application of comparative statics to equation (13) to obtain an expression for dx_{ij}/dx_{1j}, which is in turn substituted into equation (A1).

$$\text{(A2)} \quad (1+\lambda)[p^{1j} - \frac{\partial C^1}{\partial x_{1j}}] = -\lambda x_{1j}[\frac{\partial p^{1j}}{\partial x_{1j}} - a^j_{1.}[B^j]^{-1}a^j_{.1}], \quad j = 1, \ldots, n$$

Finally, we note that if the demand matrix $[A^j]$ were inverted, the upper left-hand element (q^{-1}) of the inverted matrix would be

$$q^{-1} = [\frac{\partial p^{1j}}{\partial x_{1j}} - a^j_{1.}[B^j]^{-1}a^j_{.1}]^{-1}$$

which, after some algebra, implies that

$$\text{(A3)} \quad q = \frac{p^{1j}}{x_{1j}\epsilon_p^{1j}} = \frac{\partial p^{1j}}{\partial x_{1j}} - a^j_{1}[B^j]^{-1}a^j_{.1}$$

Equations (A2) and (A3) imply that

$$\text{(A4)} \quad \frac{p^{1j} - (\partial C^1/\partial x_{1j})}{p^{1j}}\epsilon_p^{1j} = -\frac{\lambda}{1+\lambda}; \quad j = 1, \ldots, n$$

which can be restated as equation (15) in the text.

Appendix B—Further Development of Figure 3

If mode 1 can just break even as it maximizes its profit without regulation, then under partially regulated second best mode 1 would effectively have to maximize profit in order to satisfy the break even constraint. To further develop the point, under deregulation mode 1 would choose $(x_{11}, \ldots, x_{1n})$ to

$$\max_{(x_{11}, \ldots, x_{1n})} \quad \Pi^1 = \sum_{j=1}^{n} p^{1j}x_{1j} - C^1$$

At an interior optimum, the first-order conditions are of the form:

$$p^{1j} - \frac{\partial C^1}{\partial x_{1j}} = -x_{1j}(\frac{\partial p^{1j}}{\partial x_{1j}} + \sum_{i=2}^{m}\frac{\partial p^{1j}}{\partial x_{ij}}\frac{dx_{ij}}{dx_{1j}})$$

for $\quad j = 1, \ldots, n$

This equation can be rewritten using equation (A3) as

$$(p^{1j} - \frac{\partial C^1}{\partial x_{1j}})\,\epsilon_p^{1j} = -1$$

It now becomes clear that mode 1 prices will be set so that their deviations from marginal costs will be inversely related to the price elasticity of demand, for both the unregulated and partially regulated second best schemes. As the maximum profit achievable by mode 1 without regulation approaches zero, then λ becomes very large at market-clearing second best. Alternatively, $\lambda/(1+\lambda) \to 1$ in equation (A4).

REFERENCES

W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

G. H. Borts, "The Estimation of Rail Cost Functions," *Econometrica*, Jan. 1960, *28*, 108–31.

R. Braeutigam, "The Regulation of Multiproduct Firms: Decisions on Entry and Rate Structure," unpublished doctoral dissertation, Center Res. Econ. Growth, memo. no. 198, Stanford Univ., Apr. 1976.

———, "Freight Transportation: The Regulation of Intermodal Competition," in *Transportation Res. Record*, forthcoming.

G. R. **Faulhaber,** "On Subsidization: Some Observations and Tentative Conclusions," paper presented at the Office of Telecommunications Policy Research Conference on Communications Policy Research, Washington, Nov. 1972.

———, Cross Subsidization: Pricing in Public Enterprise," *Amer. Econ. Rev.*, Dec., 1975, *65*, 966–77.

A. F. **Friedlaender,** "Hedonic Costs and Economies of Scale in the Regulated Trucking Industry," in *Proc. Workshop on Motor Carrier Economic Regulation*, Apr. 1977, National Academy of Sciences, 1978.

Z. **Griliches,** "Cost Allocation in Railroad Regulation," *Bell J. Econ.*, Spring 1972, *3*, 26–41.

Donald V. Harper, *Transportation in America: Users, Carriers, Government*, Englewood Cliffs 1978.

Lawrence R. Klein, *A Textbook of Econometrics*, Englewood Cliffs 1973.

R. G. Lipsey and K. Lancaster, "The General Theory of Second Best," *Rev. Econ. Stud.*, Jan. 1956, *24*, 11–32.

H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, *60*, 693–705.

T. G. Moore, "Deregulating Surface Freight Transportation," in Almarin Phillips, ed., *Promoting Competition in Regulated Markets*, Washington 1975.

Dudley F. Pegrum, *Transportation Economics and Public Policy*, Homewood 1973.

F. P. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1972, *37*, 47–61.

R. D. Willig, "Consumer's Surplus Without Apology," *Amer. Econ. Rev.*, Sept. 1976, *66*, 589–97.

E. E. Zajac, "Some Preliminary Thoughts on Subsidization," paper presented at the Office of Telecommunications Policy Conference on Communications Policy Research, Washington, Nov. 1972.

Federal Communications Commission, "Revisions of Tariff FCC No. 260 Private Line Services, Series 5000 (TELPAK)," Docket 18128, 61 FCC 2d 606, Nov. 26, 1976.

Interstate Commerce Commission, "Rules to Govern the Assembing and Presenting of Cost Evidence," Docket 34013, 337 ICC 298, July 30, 1970.

On the Information Content of Prices

By KENNETH D. GARBADE, JAY L. POMRENZE, and WILLIAM L. SILBER*

In a recent article Sanford Grossman and Joseph Stiglitz addressed the issue of the information content of prices. They showed that in a stationary equilibrium prices may communicate information in the sense that a group of uninformed market participants will be able to infer information known to other informed participants as a function of the level of a market-clearing price. This result is of interest not only for its implications for the efficient markets literature, but also because it implies a breakdown of the classical separation between supply and demand schedules. In particular, the location of the demand schedules of uninformed transactors will depend on how much information is communicated through an observed equilibrium price, and hence will depend in part on the structure of supply or demand elsewhere in the market. Grossman and Stiglitz also showed that in some cases an equilibrium price may be a perfect aggregator of information, in the sense that it efficiently reveals all the information known by each participant. In the presence of such a perfect aggregator, the particular items of information available to any individual become redundant (see also Grossman). Although the logic of Grossman and Stiglitz is compelling, there has been no attempt to test empirically the existence or importance of their conjectures.

This paper reports the results of an empirical investigation into the information content of prices in dealer markets. Our interest here is, first, whether or not securities dealers extract information from the price quotations of other dealers and, second, whether they form their own price quotations solely on the basis of such other dealer quotes.[1] To test these propositions we specify and estimate a model of the revision of dealer quotations as a function of the observed quotations of other dealers.

Unlike the textbook example, most financial assets do not trade in discrete auction markets which lead to readily observed equilibrium prices equating supply and demand. Instead, dealers or inventory specialists quote reservation bid and offer prices at which they are prepared to transact with other dealers and public customers for their own account. The offer (bid) price of a particular dealer locates the position of his excess supply (demand) schedule and is the price at which he is willing to sell (buy) securities in transactions of a conventional size.

Dealers change their quotations from time to time as a function, in part, of their inventory position and the flow of purchase and sale orders which they observe, or more generally, as a function of new information which becomes available to them. The true equilibrium price at which a dealer might expect to see equal rates of purchases and sales is an uncertain parameter.[2] Small dis-

*Garbade and Silber are associate professor and professor, respectively, of economics and finance, Graduate School of Business Administration, New York University; Pomrenze is vice president in the government securities dealer department, Bankers Trust Company. Thomas Guba (of Paine, Weber, Jackson and Curtis) and Jay Pomerenze were the two recording Government National Mortgage Association dealers. We would like to acknowledge the valuable comments of William Baumol, Avraham Beja, and an anonymous referee on earlier versions of this paper.

[1]While our analysis is phrased in terms of dealer quotations, it is equally applicable to any other category of investor. As a data source, however, dealers are especially appropriate since they provide reservation purchase and sale quotations as part of their normal course of business.

[2]We use the term "equilibrium" price as meaning that price which would bring into equilibrium a contemporaneous Walrasian auction in which all investors participated. Dealers, like other market participants, would be unwilling to buy at a price greater than the equilibrium (conversely, to sell at a price less than the equilibrium), were that price known. Unless a dealer has a specific preference for either buying or selling (see the discussion at the beginning of Section IIIA), we assume he will bid (offer) as far below (above) his

crepancies between a quoted price and the true but unknown equilibrium price may not be very important for most investors. For dealers, however, who trade often and on highly margined positions, it is very important to locate the equilibrium price frequently during a trading day. Even a small error may wipe out the expected profit from a purchase or sale. Thus, we anticipate that dealers are particularly sensitive to extracting all available information on the location of an equilibrium price.

At any moment of time each dealer will generally know some information which is not available to his competitors, such as the price and size of trades which he has completed recently with retail customers. It is reasonable to attribute, in part, differences among the bid and offer (or ask) quotations of different dealers to such disparities in information. It follows that individual dealers may attempt to learn new information by observing the quotations of their competitors. If, for example, a dealer finds he is quoting below the average price of other dealers he may infer that he has underestimated the level of net public demand. We anticipate that if dealers acquire valuable new information by observing the reservation prices of other dealers, then they will be led to adjust their own prices as a function of their observations. This implies that the position of the supply/demand schedules of a dealer depend, in part, on the positions of his competitors' schedules.

The estimates of the equilibrium price level which are made by various dealers may differ by small or large amounts, depending on the disparity of the information available to the dealers. If the disparity is large, so that dealer quotes have a relatively large dispersion, the average of several observed prices may be a relatively poor indicator of the equilibrium price. If, on the other hand, the observed quotes are substantially similar, the mean price may be a better indicator of the equilibrium price. We conjecture that the propensity of a dealer to revise his quotes in light of the quotes of other dealers will be a decreasing function of the dispersion of the observed quotes.

The quality and reliability of the information revealed in the prices of other dealers is an inverse function of the dispersion of those prices. This aspect of the information content of prices was not present in the analysis of Grossman and Stiglitz since those authors treated price as a single number generated in a discrete auction market. George Stigler, however, has pointed out that even in a world of pure competition, as long as information on the price quotations of competing buyers and sellers is costly, we should not be surprised to find dispersion of those quotations. Garbade and Silber have shown that price dispersion exists even in the highly competitive U.S. Treasury securities market. It is, therefore, important to consider whether such dispersion affects the quality of the information content of prices.

Grossman and Stiglitz raise the question of whether equilibrium prices are perfect aggregators of information. From the point of view of the present study we inquire whether a dealer treats his own information as redundant once he has observed the reservation prices of other dealers. Since this proposition is quite strong, we also consider the same question for the special case where he observes complete agreement among the prices of other dealers. In this case the particular information known to the dealer would be considered redundant if the dealer moves his own price to the common observed price. Whether observation of other dealer prices conveys valuable information to a dealer, and whether those

estimate of the equilibrium price as possible, with the spread between bid and offer prices constrained by competition from other dealers and other market participants. See Harold Demsetz and George Benston and Robert Hagerman for a discussion of the determinants of the bid/ask spread of a single dealer, and why that spread will normally straddle the equilibrium price. Garbade and Silber consider the nature of the expected transaction spread which an investor faces in a competitive dealer market. Because of the competitive constraints on the size of a single dealer's spread, one can speak of equilibrium bid and offer prices. However, since the bid/ask spread may vary through time and across different securities, bid quotations will not be linked mechanically to offer quotations, and we consider each separately.

prices convey all available information, are empirical issues on which we present evidence below.

The remainder of the paper is divided into four sections. Section I specifies a parametric model of the revision of dealer prices as a function of the observed quotations of competing dealers. Section II describes the application of this model to a specific market: over-the-counter trading in Government National Mortgage Association (*GNMA*) pass-through securities. We discuss why this particular market was selected for the analysis. Section III presents our empirical evidence and Section IV offers some concluding remarks.

I. A Model of Dealer Price Revision

In this section we present a model of how a dealer might revise his quotations as a function of the observed quotations of other dealers. The purpose of the model is to suggest what variables are likely to be important in the revision process and to provide a specification suitable for empirical testing.[3] For simplicity of exposition we assume here that the dealer is trying to quote at a level which will lead to approximately equal rates of purchases and sales, that is, that he is seeking to quote an equilibrium price. The modifications to the basic model required when the dealer is interested primarily in being a buyer or seller are incorporated in Section III when we derive a functional form for empirical estimation.

Suppose at some point in time a dealer quotes a price p_o, which is a function of the information currently available to him. The following analysis applies equally whether p_o is a bid or an offering price. Since the dealer is seeking to execute purchases and sales with equal probability, p_o will be a noisy but unbiased estimator of the equilibrium price (bid or offer as the case may be). Let σ^2 be the normal or typical variance of the dispersion of dealer quotations, and let E be the true but unobserved equilibrium price. Since p_o is an unbiased estimator of E we can assume it comes from the model:

$$(1)\qquad \begin{aligned} p_o &= E + u_o \\ u_o &\sim N(0, K\sigma^2) \end{aligned}$$

where K is a constant which reflects the confidence which a dealer places in his own estimate of the equilibrium price relative to the normal dispersion of dealer quotes.

Now let the dealer observe a vector of prices $(p_1, \ldots, p_n)$ quoted by n of his competitors. Since the dealer will usually have no prior information on whether a particular competitor is a net buyer or seller, each of the observed quotes p_i is also an unbiased estimator of E with some estimation variance, say $F\sigma^2$. The observed prices will therefore follow the model:

$$(2)\qquad \begin{aligned} p_i &= E + u_i \qquad i = 1, \ldots, n \\ u_i &\sim N(0, F\sigma^2) \end{aligned}$$

Let P be the $n + 1$ dimensional price vector $(p_o, \ldots, p_n)$, U the $n + 1$ dimensional vector of estimation errors $(u_o, \ldots, u_n)$, and I an $n + 1$ dimensional column vector of ones. Equations (1) and (2) may be combined in the vector equation

$$(3)\qquad \begin{aligned} P &= IE + U \\ U &\sim N(0, \sigma^2\Omega) \end{aligned}$$

where $[\Omega]_{1,1} = K$, $[\Omega]_{i,i} = F$ for $i > 1$ and $[\Omega]_{i,j} = 0$ for $i \neq j$.

The Aitken estimator (see Arthur Goldberger, p. 233) of E from the model of equation (3) is

$$(4)\qquad e = (I'\Omega^{-1}I)^{-1} I'\Omega^{-1} P$$

which may be reduced to

$$(5)\qquad e = p_o + [1 + (F/nK)]^{-1}(\bar{p} - p_o)$$

where $\bar{p}$ is the mean observed quote,

[3]For simplicity we develop the model from a classical estimation viewpoint. Although similar results may be derived from a Bayesian analysis, as in Grossman, those results do not lead to simple analytic expressions and their complexity would obscure our basic points. Specifically, the prior density function of the variables E and F shown in equation (2) below is not a natural conjugate of the likelihood function of the observations $(p_1, \ldots, p_n)$ and the marginal posterior density function of E cannot be integrated analytically.

$$(6) \quad \bar{p} = n^{-1} \sum_{i=1}^{n} p_i$$

Note that the estimator in (5) is independent of the value of the normalizing variance σ^2 of the typical dispersion of observed quotations.

As we will see shortly, equation (5) forms the basis of our empirical investigation. It focuses on the estimation of the equilibrium market price as a function of the dealer's initial price and the prices he observes from other dealers. The process of dealer price revision can be made quite explicit by subtracting p_o from each side of equation (5). The magnitude of the revision $e-p_o$ is then a function of 1) the difference between the mean observed quote $\bar{p}$ and the dealer's initial price p_o, and 2) a revision coefficient $[1 + (F/nK)]^{-1}$ which depends on the number of observed dealers n, the relative dispersion of their quotes F, and the relative confidence of the dealer in his own initial quote K.

One of the determinants of the magnitude of the revision coefficient is F, the ratio of the variance of the prices quoted by other dealers around the equilibrium price to the normal variance of such observations. We assume the dealer knows the normal dispersion of quotations σ^2 for his market. Before he has observed his competitors' quotations the dealer believes those quotes will exhibit the normal variance so that F is expected to equal unity. If, however, he observes a relatively compact distribution of quotes, then he is prepared to believe that those quotes are better than usual estimators of E, i.e., that $F < 1$. Conversely, if the observed distribution is more disperse than normal then he believes $F > 1$. Note that F may vary from one set of observations to another. An unbiased estimate of F from the observed quotes is

$$(7) \quad f = \sum_{i=1}^{n} (p_i - \bar{p})^2/((n-1)\sigma^2)$$

However, since the dealer has an a priori belief that $F = 1$ he may estimate F as a weighted average of unity and f, or

$$(8) \quad \bar{f} = (w)(1) + (1-w)f \qquad 0 \leq w \leq 1$$
$$= w + (1-w)f$$

where the weighting factor w is a parameter to be determined.[4] With $\bar{f}$ given by equation (8) replacing F in equation (5), the estimate of the equilibrium price becomes

$$(9) \quad e = p_o + [1 + (w + (1-w)f)/(nK)]^{-1}(\bar{p} - p_o)$$

Equation (9) is our basic model of dealer price revision. Since, by hypothesis, the dealer is seeking to quote at the equilibrium price we expect him to move his own quote from p_o to e after observing the prices $(p_1, \ldots, p_n)$ of his competitors. The bracketed revision coefficient on $(\bar{p} - p_o)$ can be decomposed into factors on which we have empirical observations, such as f and n (the number of dealers in the sample), and the unobserved factors K and w.

From equation (5) we note that the smaller is F (the actual variance of the dispersion of dealer prices relative to its normal level), the larger will be the magnitude of the dealer's price revision for a given discrepancy between $\bar{p}$ and p_o. While the true value of F is an unobserved parameter of the revision model, the dealer forms an estimate of F based on his prior estimate of unity and on the sample of quotes from other dealers. From equation (9) it can be seen that, as long as the term w is not equal to unity, the smaller is f (the observed relative dispersion of dealer prices) the more aggressively will the dealer revise his quote towards the mean observed quote. The coefficient w will be less than unity as long as the dealer does not totally disregard the information on price dispersion contained in his sample $(p_1, \ldots, p_n)$. While we expect w to lie between 0 and 1, and not at either extreme, this is an empirical issue whose resolution will be forthcoming from our esti-

[4]Estimation of F by taking a weighted average combination of unity and f is similar to the approach suggested by Oldrich Vasicek (1973) in another context. A more precise formulation of the weighting scheme is possible from a Bayesian approach, but our data do not justify the greater complexity of such an analysis.

mated coefficients. Our discussion does have intuitive appeal in that a more compact distribution of observed prices can be said to convey higher quality information than a diffuse distribution, thereby leading the dealer to revise his quotes more aggressively.

The parameter K measures the precision of the dealer's original estimate of the equilibrium price. From equation (9) it is clear that if $K = 0$ the dealer will retain his original estimate of the equilibrium price p_o regardless of the observed quotations. Empirically, this would imply a coefficient on $(\bar{p} - p_o)$ not significantly different from zero. This result makes sense in view of equation (1), which says that if $K = 0$ the dealer believes that p_o estimates the equilibrium price without error. At the other extreme, when $K = \infty$, the coefficient on $(\bar{p} - p_o)$ is unity, irrespective of the values of f, w and n, implying complete adjustment of e to $\bar{p}$. In this case the dealer *always* subordinates his own information in favor of the information provided by the market. From equation (1) this makes sense because $K = \infty$ implies the dealer has no confidence in his initial quote.

The term w reflects the weight the dealer places on the observed measure of price dispersion f relative to his prior notion that the dispersion is equal to its normal level (see equation (8)). As mentioned above, as long as $w < 1$, f will influence the magnitude of the revision towards $\bar{p}$. If $w = 1$ the dealer places all the weight on his prior notion that $F = 1$ and the coefficient of $(\bar{p} - p_o)$ is independent of f. If $w = 0$ then f contains all the relevant dispersion information. Moreover, if $w = 0$ and $f = 0$ (so that there is no observed price dispersion), then the revision coefficient on $(\bar{p} - p_o)$ is unity and the dealer moves his quote to coincide with $\bar{p}$. This follows from the fact that $\bar{p}$ is a perfect estimate of the equilibrium price when $f = 0$ and when the dealer places no weight on his prior notion that $F = 1$.

An interesting implication of $w > 0$ is that even if the dispersion of the observed quotations is zero, the dealer will not believe that observed prices are drawn from a singular distribution because he places some weight on his prior estimate of F. Thus, observed prices will not be perceived as a perfect aggregator of information even if there is no dispersion in the observed prices ($f = 0$) as long as $w > 0$. Estimation of equation (9) will permit us to determine whether w lies at either extreme (0 or 1) or whether it takes on some intermediate value.

From the preceeding discussion we can summarize three propositions concerning the information content of dealer prices.

PROPOSITION 1: *The prices of other dealers never contain any new information, so that the coefficient on* $(\bar{p} - p_o)$ *in equation* (9) *is identically zero. This is the case of* $K = 0$.

PROPOSITION 2: *The mean observed price always contains all information, so that the revision coefficient is identically unity. This is the case of* $K = \infty$.

PROPOSITION 3: *Other dealer prices may contain some information (when* $\bar{p} - p_o$ *has a nonzero value) but the quality of that information is an increasing function of n and a decreasing function of the relative dispersion f of those prices. As a special case of this proposition, the mean price will contain all information if the revision coefficient equals unity when* $f = 0$, *i.e., when the observed prices are all identical (this is the case when* $w = 0$).

Section III presents empirical tests of the three propositions. In the next section we describe the market setting from which we collected our data.

II. Application of the Model to a Market with Incomplete Information

The choice of a market in which to gather data to test the proposition that prices contain information turns out to be nontrivial. Suppose, for example, dealer prices do contain information, but that each dealer can

continually monitor the prices of his competitors. This is the case with securities traded through the National Association of Securities Dealers Automated Quotation system (*NASDAQ*), where a montage of dealer bids and offerings is displayed on a video screen. If we asked a dealer to first write down his own quotes, next look at the quotes of his competitors, and then write down his revised quotes, we would almost always find no change between his initial and final markets. The reason, of course, is that the initial quotes already took into account the current prices of his competitors as they were disclosed on the screen. Similar comments apply to the floor market on a securities exchange, where price information is disseminated quite promptly by oral communication.

From an empirical point of view we believe it would be better to choose a market where dealers are not in continuous contact with each other, so that the process of observing other dealer quotes can be modeled by the discrete search phenomenon implied in the specification of our model. The market for Government National Mortgage Association (*GNMA*) pass-through securities meets this requirement.

The *GNMA*s are traded over the counter by about a dozen major dealer firms. The *GNMA* trader or market maker at each of these firms communicates with his retail customers and with other dealers exclusively by telephone. There is no continuous dissemination of quotes over an electronic screen or similar device. At irregular intervals during the trading day a dealer will check the markets of several of his competitors. The frequency of such observations depends upon the volatility of prices. If prices are changing rapidly a dealer will typically check his competitors more frequently to stay "in touch" with the market.

For our purposes the essential feature of the *GNMA* market is the fact that a dealer has to make an overt act to sample the prices of his competitors. We asked two dealers to monitor their normal operating procedures over the course of several weeks. First they recorded their initial bid and offer prices on a *GNMA* security, then they called several other dealers (the number varied from two to four) and wrote down the quotes of those dealers, and finally they recorded their own revised bid and offer quotations. It is important to emphasize that each dealer typically reciprocated in revealing his own quotes to his competitor, so that our data reflect true market prices. In particular, our subjects and the dealers they contacted were always liable to being "hit" on their quotes, and hence had to provide actual purchase and sale reservation prices. Every bid and offer in our data could have resulted in transactions of several hundred thousand dollars worth of *GNMA*s.

A. *Data*

The data consists of 234 sets of observations, with 117 observations on bid prices and 117 on offer prices.[5] Each set consisted of the dealer's initial quote, the observed quotes of from two to four other dealers, and the dealer's revised quote. In addition, the subject dealer recorded whether he was consciously moving his quotes away from what he believed to be the equilibrium because he wanted to be primarily a buyer or a seller (see the discussion in the next section on the need for this information).

The normal variance of the dispersion of dealer quotations σ^2 was estimated from the full data set. Let $p_i(k)$ be the quote of the ith dealer in the kth round of observations, let $\bar{p}(k)$ be the mean price in the kth round, and let $n(k)$ be the number of dealers observed in the kth round. The variance σ^2 was estimated as

$$\hat{\sigma}^2 = \sum_{k=1}^{234} \left\{ \sum_{i=1}^{n(k)} [p_i(k) - \bar{p}(k)]^2 \right\} \div \sum_{k=1}^{234} \{n(k) - 1\}$$

Using this for the value of σ^2, the value of

[5]Copies of the raw data are available from the authors upon request. The dealers did not constantly keep their bid and offer quotes in a fixed spread relationship to each other, so that each pair of quotes provide two data points rather than one.

f in each round was computed as in equation (7).

Prices were recorded in units of percent of principle value, with fractions of a percent expressed in 32ds or 64ths, as is the convention in the GNMA market. All of the results presented below are in terms of price expressed directly in percent of principle value.

III. Empirical Results

Our empirical results are divided into two parts. The first part tests the polar propositions that the mean observed quote contains no new information ($K = 0$) or that it always contains all information ($K = \infty$). We are led to reject both of these extreme cases. The second part tests the proposition that the mean price contains some information, the quality of which varies with price dispersion and the number of dealers contacted, and that it contains all information when observed prices are all identical.

A. *Test of the Polar Propositions*

Consider a simple price adjustment model derived from equation (5):

$$e = p_o + Q(\bar{p} - p_o) \tag{10}$$

If the mean observed price never carries any new information which the dealer considers relevant, then $e = p_o$ and Q should equal zero. If the mean observed price carries all information, then $e = \bar{p}$ and Q should equal unity. Thus the polar propositions may be tested simply by estimating the value of Q in equation (10).

Before testing these hypotheses on Q, however, we have to recognize that from time to time a dealer may consciously quote away from what he believes to be the equilibrium price in order to complete more sales than purchases or more purchases than sales. In either case he is trying to move his inventory to some desired level, and is willing to pay a small premium to accomplish his objectives. This means that even if $\bar{p} = p_o$, the dealer may nonetheless move his revised price above the mean price if he is trying to attract securities, or may move his price down if he is trying to sell securities. To account for desired inventory movements we asked each dealer to record, in each round of observations, the value of a shift variable S reflecting whether he was trying to buy securities ($S = +1$), or sell securities ($S = -1$), or was indifferent ($S = 0$). If the dealer is quoting his own price above what he believes is the true equilibrium by the amount γS, depending upon the particular value of S, then his prior estimate of the equilibrium is $p_o - \gamma S$. Given the sign convention on S, γ should be a positive number.[6] In this case equation (10) becomes

$$e = (p_o - \gamma S) + Q[\bar{p} - (p_o - \gamma S)]$$

After revising his estimate of the equilibrium price e the dealer will quote the revised price $r_o = e + \gamma S$, implying as the revision model:

$$r_o = p_o + Q(\bar{p} - p_o) + \gamma Q S$$

This empirical analogue to equation (10) is specified in terms of the observed revised quotation r_o rather than the unobserved estimate e, and takes into account price shifts which result from a specific inventory policy.

From the preceeding discussion we have as a function suitable for estimation, the form

$$r_o = b_1[p_o] + b_2[\bar{p} - p_o] + b_5[S] \tag{11}$$

The hypotheses are that $b_2 = 0$ if the mean price $\bar{p}$ contains no information, and that $b_2 = 1$ if $\bar{p}$ contains all information. The coefficient b_5 on S measures the extent to which a dealer will move his quotes when he is a buyer ($S = +1$) or a seller ($S = -1$); b_5 is expected to be significantly greater than zero if $b_2 > 0$. The coefficient b_1 should be unity since if ($\bar{p} - p_o$) and S are both zero

[6]The magnitude of the price shift parameter γ will depend on the preferences of the dealer for moving his inventory position quickly, and therefore eliminating what he perceives as undesirable risk exposure, vs. paying a relatively smaller premium to accomplish that objective. The magnitude will be a function, *inter alia*, of the size of dealer bid/ask spreads, the propensity of other investors to search for the best available bid and offer prices, and the degree of price dispersion in the market.

TABLE 1—ESTIMATION RESULTS[a]

Regressor Coefficient	p_o b_1	$\bar{p} - p_o$ b_2	$(\bar{p} - p_o)/n$ b_3	$f(\bar{p} - p_o)/n$ b_4	S b_5	R^2	Standard Error	Text Equation
1	1.0054 (0.0035)	0.7351 (0.0420)			0.0558 (0.0039)	0.98	0.0403	(11)
2	1.0	0.7385 (0.0421)			0.0559 (0.0039)	0.65	0.0404	(11)
3	1.0061 (0.0035)	1.0287 (0.2144)	−0.5495 (0.5844)	−0.2032 (0.0784)	0.0576 (0.0039)	0.98	0.0399	(13)
4	1.0	1.0481 (0.2150)	−0.6156 (0.5857)	−0.1915 (0.0784)	0.0577 (0.0039)	0.66	0.0401	(13)
5	1.0	1.0	−0.4895 (0.1594)	−0.1834 (0.0696)	0.0575 (0.0038)	0.57	0.0400	(13)

[a]Standard errors of estimate shown in parentheses. Coefficients without standard errors were imposed on the model, and standard errors of other coefficients in the same regressions are conditional on those maintained hypotheses.

the dealer should not revise his price, i.e., r_o should equal p_o.

The estimation results for equation (11) are shown in lines 1 and 2 of Table 1. Line 1 records an estimate of .735 for b_2. Its standard error of estimate allows us to reject the hypotheses that $b_2 = 0$ or $b_2 = 1$.

At this point we can conclude that observations on the prices of other dealers do provide a dealer with some new information, but that the mean observed price does not always aggregate perfectly all information. Thus, K in equation (9) above does not take the extreme value of zero or infinity. Line 2 of Table 1 estimates equation (11) with the constraint that $b_1 = 1$. A comparison of the first two lines indicates that b_1 is not significantly different from unity and that imposition of the constraint has no appreciable effect on the estimates of b_2 and b_5. Lines 1 and 2 also show that b_5 is significantly positive. These results for b_1 and b_5 conform to our hypotheses.

B. *Tests on the Quality of Information*

To test for the effect of the dispersion and other characteristics of observed prices on a dealer's revisions, consider a first-order expansion of equation (9) around the point $K^{-1} = 0$.

$$(12)\qquad e = p_o + [1 - K^{-1}\{(w/n) + (f(1 - w)/n)\}](\bar{p} - p_o)$$

To derive a form for estimation we replace the revised estimate of the equilibrium price e with the actual revised price r_o and, as in equation (11), add the shift variable S, obtaining

$$(13)\qquad r_o = b_1[p_o] + b_2[\bar{p} - p_o] + b_3[(\bar{p} - p_o)/n] + b_4[f(\bar{p} - p_o)/n] + b_5[S]$$

As in the previous model b_1 should equal unity. From the linearization of equation (12) we see that b_2 should also equal unity. Comparing equations (12) and (13) we can identify the coefficients $b_3 = -w/K$ and $b_4 = -(1 - w)/K$ when the constraints $b_1 = 1$ and $b_2 = 1$ are imposed. We expect that both b_3 and b_4 should be nonpositive, reflecting the impact of a larger number of quotes and the effect of less price dispersion on the revision process. The coefficient b_5 should be positive, as above.

Lines 3 through 5 of Table 1 present estimation results for several versions of equation (13), varying according to the restrictions imposed on the coefficients. Line 3 shows the unconstrained estimation results and line 4 imposes the constraint that $b_1 = 1$. The hypothesis that $b_1 = 1$ cannot be rejected and that constraint is imposed subsequently. The inventory shift parameter b_5 is significantly positive in all cases.

In lines 3 and 4 (as well as in line 5), b_4 is negative and statistically significant. This

conforms to our hypothesis that the revision of a dealer's quote towards the mean observed quote is a decreasing function of the dispersion of observed prices. The quality of the information contained in $(p_1, \ldots, p_n)$, as measured by the dispersion of those prices relative to its normal value, is a significant determinant of how far e moves towards $\bar{p}$.

The regression results of lines 3 and 4 both show that b_2 is not significantly different from its hypothesized value of unity. Line 5 reports our estimation results when the constraint $b_2 = 1$ is imposed.[7] Here we find that both b_3 and b_4 are negative and statistically significant, as expected. From the identification of $b_3 = -w/K$ and $b_4 = -(1 - w)/K$ we can reject the hypothesis that w equals either zero or unity and therefore conclude that it lies between those extreme values. From the regression results of line 5 we can identify $w = .73$ and $K^{-1} = .6729$, or $K = 1.48$. The fact that w is greater than zero indicates that even if the observed dispersion is zero, the dealer does not revise his quote completely towards $\bar{p}$ since he does not ignore his prior estimate that the observations $(p_1, \ldots, p_n)$ come from a population with some positive variance. Thus, even if the observed quotes are identical, a dealer will not treat $\bar{p}$ as a perfect aggregator of information. Similarly, the fact that K is neither 0 nor infinitely large confirms the results reported in lines 1 and 2 of our table (and suggested by equation (1) above), that the dealer does not completely ignore the information conveyed by $\bar{p}$, nor does he totally subordinate his own information to that which he derives from searching the market.[8]

The empirical observation that dealers change their quotations as a function of their competitors' prices could also be explained by an analysis which does not rely on the revision of prices due to acquisition of new information. Specifically, a dealer may offer (bid) at as high (low) a price as is compatible with his competitors' quotations. When other dealers are quoting offer prices substantially above his offer, he would move his to just below their level. This "competitive" explanation would imply that a dealer is interested in quotes of other dealers only to the extent that he wants to avoid being priced out of the market.

We have two reasons for believing that prices in our model are revised according to the information process described in Section I rather than because of this exclusively competitive explanation. First, the price adjustment scheme implied by the latter analysis is not symmetric. Thus, a dealer will raise his offering price (but not his bid), or lower his bid price (but not his offer), in response to the prices of other dealers. This implies that a dealer will widen his bid/ask spread if observation of his competitors' prices leads him to revise either his bid or his offer price. However, in only 7 percent of our observations where a dealer revised at least one of his prices did he widen his spread in response to observing his competitors' prices.

Second, consider again the estimation results for the simple model of equation (11) reported in line 2 of Table 1:

$$r_o = p_o + \underset{(.0421)}{.7385}\,(\bar{p} - p_o) + \underset{(.0039)}{.0559}\,S \tag{14}$$

[7]Several additional constrained versions of the model were also considered. In particular, while line 4 suggests that b_2 may equal unity and b_3 equal zero simultaneously, that joint null hypothesis was rejected by an F test at a 1 percent level of confidence. Our model suggests that the constraint $b_2 = 1$ is appropriate, and the result of that constraint is reported in line 5 of Table 1.

[8]All of the regression results reported in Table 1 were tested for structural stability across several subdivisions of the full data set. These tests included whether the coefficients of the model were the same depending on whether the dealer was a buyer, seller, or neutral; whether he was a buyer or seller vs. whether he was neutral; and whether he was a buyer vs. whether he was a seller. We also tested whether the two recording dealers had the same behavioral parameters. Using an F test we were unable to reject at a 5 percent confidence level the hypothesis that our results were invariant with respect to the buyer/seller/neutral classifications. More importantly, we could not reject the hypothesis that the two dealers had the same behavioral parameters. In view of our analytical objectives, as well as the consistency of the qualitative results across subgroupings of the data, Table 1 reports the estimation results only for the full data set.

This equation shows quite clearly that the revised quotation r_o is not independent of the dealer's initial quotation p_o. If it were, the coefficient on $\bar{p} - p_o$ would not be significantly different from unity. The competitive analysis cited above, however, gives no reason for the significance of p_o to the revised quotation. If that analysis were valid we would expect to find dealer quotes following a model like:

$$r_o = \bar{p} + b_5 S \tag{15}$$

Thus, they would be relatively better bidders (and poorer offerors) when $S > 0$ and they want to buy, and relatively better offerors (and poorer bidders) when $S < 0$ and they want to sell. Similar comments apply to the results in line 5 of Table 1, which can be written

$$r_o = \bar{p} - \underset{(.1594)}{.4895}[(\bar{p} - p_o)/n] - \underset{(.0696)}{.1834}[f(\bar{p} - p_o)/n] + \underset{(.0038)}{.0575}\, S \tag{16}$$

The coefficients on $(\bar{p} - p_o)/n$ and $f(\bar{p} - p_o)/n$ have an expected value of zero under the null hypothesis of the exclusively competitive analysis.

IV. Conclusions

This paper presented empirical tests of the proposition that prices contain information. Our particular application considered the special case of whether dealers acquire valuable information from observation of the reservation purchase and sale prices of their competitors, and whether they are led to change their own quotations as a function of those observations. To test the hypothesis we developed a unique data base from a market where dealers cannot continuously monitor the changing quotations of their competitors, so that the process of acquiring new information may be modelled as a discrete search phenomenon.

Our results lead us to reject the hypotheses that observed prices convey no information and that the mean observed price contains all information. Dealers do acquire new information from their competitors, but they do not consistently treat their own information as redundant after obtaining their competitors' prices. We are also lead to reject the hypothesis that the mean observed price conveys all information even in the special case where all observations are identical.

The quality of the information carried in a particular constellation of observed prices depends on the dispersion of those prices. A more compact distribution of observations will lead a dealer to revise his estimate of the equilibrium price further towards the mean observed price. Thus, not only do prices contain information, but dealers are alert to the quality of that information.

REFERENCES

G. Benston and R. Hagerman, "Determinants of Bid-Ask Spreads in the Over-the-Counter Market," *J. Finan. Econ.*, Dec. 1974, *1*, 353–64.

H. Demsetz, "The Cost of Transacting," *Quart. J. Econ.*, Feb. 1968, *82*, 33–53.

K. Garbade and W. Silber, "Price Dispersion in the Government Securities Market," *J. Polit. Econ.*, Aug. 1976, *84*, 721–40.

Arthur Goldberger, *Econometric Theory*, New York 1964.

S. Grossman, "On the Efficiency of Competitive Stock Markets where Trades have Diverse Information," *J. Finance*, May 1976, *31*, 573–85.

——— **and J. Stiglitz,** "Information and Competitive Price Systems," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 246–53.

G. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–25.

O. Vasicek, "A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas," *J. Finance*, Dec. 1973, *28*, 1233–39.

An Essay on Monopoly Power and Stable Price Policy

By S. Y. Wu*

It has long been observed that firms with monopoly power prefer a stable price policy even under the expectation that demand will fluctuate over time. Because a different emphasis is placed on the firm's objectives and concerns, there exists a variety of explanations for the stable price phenomenon. The kinked demand theory (see Paul Sweezy), the price leadership analyses (see Gardiner C. Means, 1954, and J. Fred Weston), and the various oligopoly models using game theoretics (see Martin Shubik) all stress the presence of conjectural variation among rivals and the fears that price competition may become mutually ruinous. Not surprisingly, they conclude that a price, once established, becomes inviolable. Others (see Roy F. Harrod) argue that when a firm strives to maximize long-run profit, it will equate long-run marginal revenue to long-run marginal cost; price becomes invariant with changes in short-run demand. Still others (see Edward Mason) emphasize that when a firm produces many commodities and faces the problem of jointly setting prices for all of its products, there exist costs associated with price adjustment. These costs induce a stable price policy even if the demand for individual products fluctuates. Finally, rather than pursuing a simple goal of maximizing profits, a firm may have many objectives, for example, increasing market share, achieving target return on investment, stabilizing profit margins or meeting competition. The decision maker may set different priorities among these competing goals, and a low priority on price policy also explains stable prices over time (see Mason; Robert Lanzillotti; Herbert Simon; Weston). These analyses have undoubtedly provided explanations for stable prices under special circumstances. However, the power of these theories is limited by the narrowness of their scope, the informality of their arguments, and their reliance on motives other than profit maximization. By using the recent developments in the concepts of uncertainty and risk aversion, and in dynamic analyses, I will present a general stable price model for a profit-maximizing firm which possesses some degree of monopoly power and makes long-run decisions under the conditions of uncertainty. As shall be shown, the many seemingly divergent explanations in the stable price phenomenon are also reconciled under the proposed framework.

My departure from tradition rests crucially upon a difference in opinion concerning the nature of uncertainty faced by the firm. The prevailing view is that uncertainty is exogenously determined. In particular, decisions made prior to the realization of a random event leave unaffected the degree of uncertainty faced by the firm. Under this premise, it is natural to conclude that *ex ante* decisions always yield suboptimal results, and that if possible, all decisions should be made after the realization of the random event, that is, after all uncertainty vanishes. Moreover, decisions should be made in sequence, each depending upon information that is currently available but is not known in advance. Following this line of reasoning, Hayne Leland (p. 280) concludes that the monopolist's expected utility of profit diminishes as more of its decision variables are selected prior to the realization of the random event.

In contrast to the prevailing view, let us observe that there are two sources of un-

*University of Iowa. I wish to thank K. Brooks, K. Currie, R. Nelson, A. M. Spence; M. Balch, G. Fethke, and my other colleagues at Iowa; a referee, and especially G. Borts for helpful advice and comments.

certainty: (i) that generated exogenously and viewed by the decision maker as the state of the world, and (ii) that generated endogenously by the decision maker's attempt to exercise some conditional influence over his environment or from the interrelated nature of human behavior among the market participants. Based upon (ii) we can argue that when a firm enjoys some degree of monopoly power it not only can set its prices but also can affect the degree of uncertainty faced by the firm through its own action. Especially because the firm's *ex ante* decisions affect its *ex post* market conditions, the entrepreneurial role not only includes the traditional *technical management*, that is, the procurement of inputs and marketing of outputs, but also the *management of uncertainty*. This latter role includes: (i) the selection of uncertainty related variables to be included in the decisions and the employment of means to internalize uncertainties faced by the firm; and (ii) the determination of which variables should be selected *ex ante* and which *ex post*. It is the management of uncertainty in the latter sense that contributes crucially to the understanding of the rationality behind the selection of a stable price policy.

This paper is divided into two sections. Section I demonstrates how a stable price policy emerges through the management of uncertainty. Section II examines both theoretical and policy implications of stable prices.

I. Toward a Stable Price Policy

In order to facilitate the examination of the rationale leading to the adoption of a stable price policy, it is convenient at the outset to posit a concrete model for the firm. Assume the entrepreneur knows his production and inventory costs but does not know his demand in each period with certainty. Let the entrepreneur strive to maximize his expected utility of profit for a (T + 1) period decision horizon. His decision problem is to choose a set of policies, i.e., the paths of $\{P_t, S_t, Q_t, I_t\}$, t = 1, . . . , T + 1, so as to

$$(1)\quad \max \sum_{t=1}^{T+1} \int_\Omega \rho^{t-1} \cdot u[\pi_t(\boldsymbol{d}_t, \boldsymbol{d}_t^c, \omega_t)]dF(\omega_t; \boldsymbol{d}_t)$$

subject to

$$(C1)\quad \pi_t = P_tS_t - C(Q_t) - \Phi(I_t, \boldsymbol{d}_t)$$

$$(C2)\quad D_t = \xi_t(P_t, \beta_{t-1}, \omega_t; \boldsymbol{d}_t)$$

$$(C3)\quad S_t = Q_{t-1} + I_{t-1} - I_t$$

$$(C4)\quad S_t = \begin{cases} D_t \text{ if } D_t \leq Q_{t-1} + I_{t-1} \\ Q_{t-1} + I_{t-1} \text{ if } D_t > Q_{t-1} + I_{t-1} \end{cases}$$

$$(C5)\quad S_t, Q_t, I_t, P_t \geq 0 \text{ for } t = 1, \ldots, T+1$$

where ξ_t is the firm's demand function in period t; u is the entrepreneur's utility function; and ρ is a discount factor; P_t, π_t, S_t, Q_t, I_t, respectively, are the price, profit, sales, production, and inventory in period t; C and Φ denote the production and inventory costs; $\beta_{t-1} = g(D_{t-1} - S_{t-1})$, with $g(0) = 0$, is a goodwill penalty index, it increases as the amount of shortages in period t − 1 increases, and we assume that $\partial\xi_t/\partial\beta_{t-1} < 0$, which reflects the adverse effect of a shortage; $\omega_t \in \Omega$ is a random shift parameter underlying the firm's demand function in period t with a probability distribution function $F(\omega_t; \boldsymbol{d}_t)$; finally, $\boldsymbol{d}_t$ and $\boldsymbol{d}_t^c$ are, respectively, a set of *ex ante* and *ex post* decision variables. They are subsets of $x_t = \{P_t, S_t, Q_t, I_t\}$, where $\boldsymbol{d}_t \cup \boldsymbol{d}_t^c = x_t$ and $\boldsymbol{d}_t \cap \boldsymbol{d}_t^c = \phi$.

In the context of uncertainty management, there are a number of variations on the basic model represented by (1). A specific model will emerge depending upon the manner in which information is utilized with the passage of time, and upon which variables are determined *ex ante* and which are determined *ex post*. It will be shown that the stable price model is a particular model belonging to this family. Since the main objective of this section is to explore the rationale leading to the adoption of a stable price policy, I will also present a sequence of models belonging to (1) in a manner which will not only describe the structure of the decision problem and the anatomy of price formation but also will

help to justify a stable price policy under a maximization framework.

The main thread of the argument is that due to interaction between a seller and his buyers and among sellers, shifting price determination from *ex post* to *ex ante* and then to a stable price policy may in each case improve the distribution function underlying the seller's demand,[1] and consequently raises the (risk-averse) entrepreneur's expected utility of his long-run profit. Accordingly, this section is organized in three corresponding subsections.

A. *Prices Determined Ex Post*

In order to begin the reasoning process leading to the adoption of a stable price policy, let us start with a model where price is determined *ex post*. Frequently, due to technological or institutional limitations, it is impossible to make all decisions *ex post*. The existence of a production lag is a case in point. The presence of a production lag requires that input decisions precede production by at least the number of periods equal to the production lag. Therefore, production decisions must be made *ex ante*. Since the path of demand is not known with certainty, the *ex ante* decision on production may produce adverse consequences *ex post*. In order to mitigate the impact of this possible "erroneous decision," precautionary inventory could be set aside to cushion the random shock. Moreover, in a multiperiod decision context, inventories are accumulated not only as buffer stocks but also to serve transactions and speculative purposes. Thus, although inventory appears to be wasteful under a certainty environment, as Richard M. Cyert and James G. March have observed, it performs a productive role under uncertainty.

In terms of the model represented by (1), the decision problem faced by the entrepreneur is to select $\{Q_t\}$ *ex ante* and $\{S_t, P_t, I_t\}$ *ex post* for $t = 1, \ldots, T+1$ so as to maximize his expected utility of profit. When price is determined *ex post*, it can be adjusted freely to clear the market. In addition, information can be extracted and used to make decisions for the future and to determine the *ex ante* decision variables for the next period. Decisions are thus made in stages where each decision relies on information gained from the past. Formally, let $d_t = \{Q_t\}$ and $d_t^c = \{S_t, P_t, I_t\}$. This decision problem can best be described by and solved as a dynamic programming problem. Let $J_t(Q_{t-1}, I_{t-1})$ denote the entrepreneur's maximum expected utility of profit in period t when decisions from period t onward are made optimally. The fundamental recursive equation for the dynamic programming problem can be written as

$$(2)\quad J_t(Q_{t-1}, I_{t-1}) = \max_{Q_t} \int_\Omega \max_{S_t, P_t, I_t} \cdot \{u[\pi_t(d_t, d_t^c, \omega_t)] + \rho J_{t+1}(Q_t, I_t)\} dF(\omega_t; d_t)$$
$$J_{T+1}(Q_T, I_T) = 0 \qquad \text{for } t = 1, \ldots, T$$

subject to conditions 1 through 5 in equation (1). Note that since price can be adjusted freely to clear the market, the constraints (C1)–(C5) are modified accordingly to reflect $S_t = D_t$ and $\beta_{t-1} = 0$ for all t.

This problem is well defined. Given the initial production Q_0 and inventory I_0, the firm's optimal policies $\{Q_1^o, \ldots, Q_{T+1}^o\}$, $\{S_1^o, \ldots, S_{T+1}^o\}$, $\{P_1^o, \ldots, P_{T+1}^o\}$, and $\{I_1^o, \ldots, I_{T+1}^o\}$ can be found. Suppose that uncertainty is exogenously given and that the cost of obtaining information is negligible. In the presence of a production lag, because all information to date has been utilized, policies thus selected would necessarily be optimal. This conclusion, however, is invalid if the information and search costs are high and if the firm, by its own action, can affect its uncertain environment.

[1]The distribution G is improved compared with the distribution F if

(a) $\int_a^h [f(\omega) - g(\omega)] d\omega > 0$ for all $h \in [a, b]$

or

(b) $\int_a^h [F(\omega) - G(\omega)] d\omega > 0$ for all $h \in [a, b]$

In other words, the "improvement" may be in the form of first- or second-degree stochastic dominance. (See Josef Hadar and William Russell.)

B. *Price Determined Ex Ante*

In general, search costs increase sharply if the time available for search shortens. Determining price *ex ante* allows the seller to provide his buyers with sufficient lead time to search, and thereby reduces their search costs. This reduction in costs may ease the buyers' credit constraints and thereby translates into a higher demand for the seller's product. The validity of the above argument rests crucially upon the assumption that buyers face credit constraints under uncertainty. This assumption is intuitively plausible in the case of a consumer (see the author, 1974). As for the firm, its plausibility is based upon the observation (see Michael Balch and the author, 1974) that if risk-averse creditors do not know a firm's future profit stream with certainty, they may refuse to lend precisely at a moment when credit is critically needed for survival. The presence of an externally imposed "potential credit constraint" leads the firm to self-impose a limitation on borrowing so as not to jeopardize its borrowing power in time of need. A reduction in search costs improves the firm's profit flow and thus helps to ease its credit constraint.

Furthermore, there often exists a significant amount of endogenously generated uncertainty, and an entrepreneur with some monopoly power frequently finds that his *ex ante* decision directly affects the firm's *ex post* market conditions. Switching the price decision from *ex post* to *ex ante* may improve the distribution F. To demonstrate, we must show the interrelatedness of policies among the sellers and between a seller and his buyers.

1. *Interaction Among Sellers*

Because the sources of uncertainty can be both exogenous and endogenous, the random variable that determines a seller's demand in each period can be decomposed into two parts: (i) the part that determines the market demand; and (ii) the part that determines the individual seller's market share. Cutthroat competition arises when, with a given market demand, sellers compete for the market share through price incentives. It contributes to endogenously generated uncertainty.

Price competition can be successful only in a situation where rivals do not have the opportunity to retaliate with reciprocal price cuts. The shorter the time interval between the price quotation and the actual exchange, the smaller is the chance for a successful retaliation. When price is established after the realization of the random market demand, sales can follow immediately upon the price quotation, and retaliation may be precluded. This possible gain from price cutting often serves as an inducement for the firms to engage in price competition despite the fact that a firm's expected demand or expected market share is not appreciably influenced by price warfare; rather each is determined by factors such as relative production cost, production capacity, and inventory availability, etc. When opportunity to retaliate is present, cutthroat competition merely generates greater endogenous uncertainty. It yields for the firm a probability distribution of the market share characterized by a mean-preserving increase in risk and a probability distribution of profit characterized by a mean-decreasing increase in risk. Assuming that the utility function is concave in the random variable ω, these results will necessarily lead to a decline in the entrepreneur's expected utility of profit. Thus stated, price competition is never rational. Since opportunities for short-run gain nonetheless do exist, some will occasionally venture a gain by initiating a price cut, a price war may ensue causing a reduction in the expected utility of profit for all.

In contrast, when price policy is made and announced before the realization of the random market demand—because no sales can take place and because the rival's opportunities to retaliate abound—the price cutter knows for certain that his action can never be successful because the rivals will not ignore a price cut; consequently, price competition becomes totally ineffectual as

a means to obtain a larger market share. The entrepreneur thus will discard it as a competitive instrument. Shifting the determination of price from *ex post* to *ex ante* will also have the effect of removing the temptation for rivals to engage in price warfare and eliminate a portion of the endogenously generated uncertainties. Assuming the entrepreneur is risk averse, he will choose *ex ante* determination of prices for it is associated with a more favorable distribution function F, and yields for him a higher expected utility of profits.

2. *Interaction Between a Seller and His Buyers*

In addition to conjectural variation in demand among sellers, interaction between a seller and his buyers may also militate in favor of choosing prices *ex ante*. A seller who makes his price decision *ex ante* and announces his decision to the buyers will often receive feedback indicating the buyers' intentions to purchase. Such feedback is informative and can lead to an improvement in the seller's subjective distribution of the random variable and to an increase in his expected utility of profit. In addition, the act of announcing prices *ex ante* reduces not only the buyers' search cost but also their price uncertainty. We have seen how a reduction in the buyers' search cost increases the seller's demand; a reduction in price uncertainty for the buyers also acts in several ways to increase the seller's demand, and therefore his profit. A risk-averse buyer, who may be either a consumer or a producer using the seller's product as an intermediate input, is willing to buy more at a given price or pay a higher price for a given quantity in exchange for a reduction in price uncertainty.[2] The seller's price announcement will also eliminate the necessity for the buyers to accumulate precautionary inventory used to guard against price uncertainty.[3] Assuming that the product in question is not an inferior good to the buyer, the reduction in inventory cost will again ease the buyer's credit constraint and may be translated into higher demand for the seller's product.

In summary, *ex ante* determination of prices may be favored because shifting price setting from *ex post* to *ex ante* increases the seller's demand and improves the distribution function of the random variable underlying the firm's demand function. These results contribute to a rise in the seller's expected utility of profit.

Having provided a rationale for the firm to choose prices *ex ante*, the entrepreneur's decision problem now becomes: to select the *ex ante* decision variables $d_t^* = \{Q_t, P_t\}$ and the *ex post* decision variables $d_t^{*c} = \{S_t, I_t\}$ so that (1) is maximized subject to the constraints (C1)-(C5). The corresponding dynamic programming recursive equation now becomes

$$(3)\quad J_t^*(Q_{t-1}, I_{t-1}, \beta_{t-1}) = \max_{Q_t, P_t} \int_\Omega \max_{S_t, I_t} \{u(\pi_t(d_t^*, d_t^{*c}, \omega_t)) + \rho J_{t+1}^*(Q_t, I_t, \beta_t)\} dF(\omega_t; d_t^*) \quad \text{for } t = 1, \ldots, T$$

$$J_{T+1}^*(Q_T, I_T, \beta_T) = 0$$

subject to (C1) (C5).

Given Q_0 and I_0, (3) again yields a set of

[2]This statement cannot be made without some qualification. It has been shown by E. K. Choi and the author that when the individual's demand curve is convex to the origin, price gamble will yield for the buyer an expected gain in consumption. If this gain exceeds the buyer's risk premium, he will accept a price gamble despite the fact that he will never accept a quantity gamble. Whether this observation invalidates the statement in the text, we cannot say. It depends upon the number of convex demand curves in the market and the degree of risk averseness of the buyers— a condition that may vary among industries.

[3]When the buyer is offered a stable price over time, uncertainty faced by him is not eliminated altogether: he exchanges a price uncertainty for a quantity uncertainty. In light of the remarks made in fn. 2, it is not clear from the buyer's point of view whether this exchange is indeed desirable. It has been observed that some producers, besides maintaining a high level of inventory, have adopted a policy of announcing the production schedule in advance. (See Joseph Orlicky.) This development reflects the seller's attempt to help reduce the quantity uncertainty faced by the buyers and thus makes the stable price policy more attractive to them.

optimal policies $\{P_1^o, \ldots, P_{T+1}^o\}^*$, $\{Q_1^o, \ldots, Q_{T+1}^o\}^*$, $\{S_1^o, \ldots, S_{T+1}^o\}^*$, and $\{I_1^o, \ldots, I_{T+1}^o\}^*$. In this case, precautionary inventory serves as an equilibrator *ex post*; it will adjust to absorb the shocks generated from the random variations in the firm's demand.

C. *Stable Price Policy*

The above analysis implies that it may be profitable for the seller to determine price *ex ante*. Questions remaining are: (i) How far ahead of the realization of the random demand in a given period should the seller make his price decision? (ii) Should price be set invariant over time or change according to a predetermined and announced schedule? If the buyers' search costs decrease and the seller's profit increases with lead time in price announcement, then the longer-period price schedule will prevail. Once the firm chooses to announce the price with sufficient lead time, it is inevitable that it will also choose a single price invariant over time. First of all, price scheduling is impractical unless all competitors' price schedules over time are identical. Assuming the contrary, buyers will at each time purchase from that firm whose price is lowest, thus forcing the high price firms to revise their prices. This defeats the intended schedule. Sellers now face the alternatives of either charging the prices determined by the lower bound of all schedules or following a uniform stable price policy. Following the lower bound is obviously suboptimal, so it is not surprising to find that firms will choose the stable price policy. In addition, stable price policy may in any event be advantageous over price scheduling. Holding price invariant over time further reduces the buyers' search costs and eliminates the need to maintain speculative inventory; the resultant saving may be translated into increased demand for the seller's product. Maintenance of a stable price policy over time is not without cost for the seller since his ability to smooth production over time is reduced, and he must maintain additional inventory. However, if the seller's increased revenue due to price stability outweighs the concomitant cost increase, then it is to his advantage to adopt such a policy. This is indeed often true since economies of scale for inventory and storage exist, and the seller often enjoys a comparative cost advantage in storing his own product.

In addition to the foregoing reasons, the game-theoretic considerations reinforce the stable price policy. Specifically, if any of the competitors adopts a stable price policy, then the buyers' preferences for a stable price will compel the remaining rivals to follow suit.

Now the firm's *ex ante* decision variables are $d_t^{**} = \{Q_t, P\}$ and its *ex post* decision variables are $d_t^{**c} = d_t^{*c} = \{S_t, I_t\}$. To solve the firm's maximization problem, we rely on the following dynamic programming recursive equations:

$$(4)\quad \begin{aligned} J_t^{**}(Q_{t-1}, I_{t-1}, \beta_{t-1}; P) &= \underset{Q_t}{\text{Max}} \int_\Omega \underset{S_t, I_t}{\max}\{u(\pi_t(d_t^{**}, d_t^{*c}, \omega_t)) \\ &\quad + \rho J_{t+1}^{**}(Q_t, I_t, \beta_t; P)\} dF(\omega_t; d_t^{**}) \\ &\qquad \text{for } t = 2, 3, \ldots, T \\ J_{T+1}^{**}(Q_T, I_T, \beta_T; P) &= 0 \\ J_1^{**}(q_0, I_0, \beta_0) &= \underset{P}{\text{Max}} J_1^{**}(Q_0, I_0, \beta_0, P) \end{aligned}$$

subject to (C1)–(C5).

Given Q_0 and I_0, (4) yields a set of optimal policies P^o, $\{Q_1^o, \ldots, Q_{T+1}^o\}^{**}$, $\{S_1^o, \ldots, S_{T+1}^o\}^{**}$, and $\{I_1^o, \ldots, I_{T+1}^o\}^{**}$. Here (precautionary) inventory again serves to adsorb the random shocks *ex post*.

I have now completed my presentation of the anatomy of price formation, that is, the development of the rationale leading to the adoption of a stable price policy. The basic underlying assumption of the analysis is that the monopolist not only can set his own price but also, through his own action, can exercise some conditioning influence on his own uncertain environment. Under this circumstance, the structure of the decision problem is fundamentally different from those in the existing literature. In contrast to the *ex ante* and *ex post* two-stage decision problem posed by Leland, decisions now can be characterized as a three-stage maximization problem. During the time of

planning, the entrepreneur chooses a strategy among the various available strategies, that is, he compares the long-run expected utility of profits derived from various possible divisions of *ex ante* and *ex post* decision variables and chooses that division which gives him the highest expected utility of profits. Formally, he chooses to maximize

(5)
$$\{J_1(Q_0, I_0), J_1^*(Q_0, I_0, \beta_0), J_1^{**}(Q_0, I_0, \beta_0)\}$$

subject to the appropriate constraints. Settled on a strategy, as the market evolves from period to period, the entrepreneur faces the second-stage decision problem: in each period, he utilizes information available to date to select the values of the *ex ante* decision variables that enable him to maximize the present value of his expected utility of profit. Once the random variable in a period is realized, the entrepreneur makes the third-stage decision: he chooses the *ex post* variables that maximize the present value of the realizable expected utility of profit. This process repeats itself from period to period until the end of the decision horizon is reached.

Suppose the monopolist should choose the strategy represented by (4). Based upon this analysis of price formation, I suggest that it is endogenous uncertainty, risk aversion, and opportunity cost foregone by following a flexible price policy (rather than the more commonly emphasized direct cost associated with price adjustment) that leads the firm to adopt a stable price policy. The reader is cautioned not to interpret this view so rigidly as to have advocated the universality of the stable price phenomenon. Conditions in some markets are clearly not conducive to a stable price policy and thus may lead the entrepreneur to choose a different strategy. It has been suggested by A. Michael Spence and by the author (1975) that markets favorable to stable price policy often have the following characteristics: (i) low inventory cost, (ii) price elasticity that varies directly with the level of the firm's demand, and (iii) the firm's demand function exhibits greater random variation relative to the changes in its amplitude over time. Since this paper is concerned with the stable price phenomenon, let us now discuss the economic implications of stable prices.

II. Implications of Stable Prices

It has been shown that when a risk-averse entrepreneur is faced with an uncertain environment over which he has some conditional influence, a stable price policy may emerge as a rational long-run choice. It is now possible to examine some theoretical and empirical implications of stable prices. In the following, I shall discuss only four general topics. The first is concerned with the nature of the entrepreneur's decision process and the second, the concept of equilibrium. The last two deal with equilibrium in the market and in the economy.

A. *Decision Process*

In Section I it was observed that the structure of the firm's decision problem under the proposed framework is different from that under the traditional one. Here, it is further observed that when the price is set for the entire decision horizon, the structure of the decision problem undergoes additional change.

In passing, note that the problems represented by equations (2) and (3) can be solved in a manner consistent with the conventional decision process: Information derived from successful realization of the random variable is used to forecast the firm's future demands, and these forecasts provide the basis for the selection of the values of the decision variables in future periods. In this framework, every decision is by its nature the best that can be made. The process involved in solving equation (4) is, however, different; it can be described briefly as follows: The entrepreneur uses market and technical information available at the time of planning to forecast demands, and then selects a price policy for the entire decision horizon. Only production, sales, and inventory are determined period by period in the

usual manner; the price policy is invariant until changes in market conditions warrant a revision.

The crucial task is to detect when a change in market condition has taken place. Since the firm does not know its demand in each period with certainty, realized sales S_t and hence inventory I_t may differ from the planned optima, S_t^o and I_t^o. This difference between expectation and realization may stem from either random variations or systematic changes in market conditions. In the former case, changes in precautionary inventory will absorb the difference. In the latter case, the selected policy ceases to be optimal, and a new one must be found. A criterion is, therefore, needed to distinguish these two cases in order to determine whether the chosen policy is indeed optimal or warrants a change.

Let $I_t^* = I_t + \tau_\alpha \hat{\sigma}_I$ and $I_t^{**} = I_t - \tau_\alpha \hat{\sigma}_I$, where τ_α is the t-statistic associated with the probability α (α is the largest tolerable level of error in case the null hypothesis is rejected when it is true) and $\hat{\sigma}_I$ is the estimate of the standard deviation of inventory variations. A naive rule such as the following one may serve as a testing criterion: *Accept the null hypothesis when* $I_t^{**} \leqq I_t^o \leqq I_t^*$; *reject the null hypothesis otherwise.* The interpretation of this decision rule is: if the planned inventory for period t falls between the upper and the lower limits, although the planned and the realized inventory differ, this difference can be attributed to random causes. The entrepreneur will conclude that forecasts regarding the firm's demand are correct so far, and he will uphold his selected policies. Only when the planned inventory for period t falls outside the limits will the entrepreneur conclude that he has misjudged the market conditions, and thus should reappraise his expectations and change his policies. Moreover, the planned inventory, although not falling outside the bounds, may fall consistently near either of these bounds. In such a case, the entrepreneur may also be forced to revise his policies. He will do so either because he views this phenomenon as an indication that the firm's demand has changed systematically, or because inventory has become so large as to be financially too burdensome to maintain or so small that it becomes dangerously inadequate to absorb the random demand variations.

Unlike the conventional decision process, my analysis suggests that after the entrepreneur has forecasted demands, selected a stable price strategy and determined the relevant policies, as the events unfold he must also test the hypothesis as to whether the selected policies are indeed optimal. Thus, unlike the conventional decision process which requires continuous forecasting, the proposed process necessitates continuous testing. Continuous testing leaves the entrepreneur's decisions relatively insensitive to daily market variations and provides the firm and the market with a much desired stability.

This price insensitivity to daily market changes is often interpreted by some as reflecting the fact that entrepreneurs are not maximizers. It has been our concerted effort to show that a stable price policy can indeed be consistent with the maximization of the entrepreneur's expected utility of profit. In light of this analysis, I feel that the non-profit-maximization motive may have been mistakenly attributed to the monopolist. The statement that entrepreneurs are not maximizers but merely satisficers is a result of using a neoclassical yardstick to evaluate the firm's behavior under an uncertain environment; it must be scrutinized further with greater care.

B. *Concept of Market Equilibrium*

The traditional concept of market equilibrium reflects a state of contentment associated with a market clearance, that is, what sellers want to sell they actually sell and what buyers want to buy they actually buy. My analysis suggests that under uncertainty the optimal choice of the firm is an *ex ante* concept. Although the amount purchased equals the amount sold in each market transaction, this equality does not indicate a market clearance in the traditional sense; there exist buyers and/or

sellers whose expectations are not realized. Because both parties under uncertainty utilize precautionary inventory to absorb random shocks, observe that even though the market is not cleared *ex post*, the decision makers may nonetheless find the results satisfactory. Market equilibrium under this situation is characterized by the invariance of the established policies, not by market clearance. Thus, even if the *ex ante* decision made by the buyers and sellers does not lead to an *ex post* market clearance situation, as long as the decision makers through the testing of hypotheses uphold their established policies, the market is in equilibrium. Similar notions of equilibrium have been articulated in different situations by Roy Radner and by Leonid Hurwicz, Radner, and Stanley Reiter; its emergence in the present context further demonstrates that it is a general and useful concept.

C. *Market Equilibrium*

The model developed in Section I now can be used to investigate how all firms' policies are made consistent with each other or how an equilibrium is established in an oligopolistic market. Two problems will be examined. The first relates to the establishment of a uniform price when firms produce a homogeneous product, and the second deals with the phenomenon of market segmentation, that is, the coexistence of large and small firms in the same industry.

1. *Market Price and Market Equilibrium*

In order to answer the question "How is a common market price determined?," we must first recall an earlier observation that cutthroat price competition is strictly a short-run phenomenon and is inconsistent with the objective of long-run maximization. When oligopolists rule out the use of price as a competitive instrument and adopt a stable price policy, the establishment of market price will naturally emerge as a consequence of the individual firm's adjustment guided by its own decision rule. If the prices of the oligopolists are not the same, the inventory of the high-priced oligopolist will exceed the upper limit prescribed by his decision rule, while the inventory of the low-priced oligopolist will fall below the lower limit prescribed by his decision rule. The former will revise his price downward, while the latter will revise his upward until a common price is established. Although a uniform price is established for all traders in an oligopoly market, this price will differ from that which would have been established if the market were competitive. In a competitive market, the seller balances expected price to marginal cost plus risk premium, while in an oligopoly market it can be shown formally on the basis of (4) that the seller balances expected marginal revenue to marginal cost plus risk premium—an act which reflects the seller's market power.

Market price may converge potentially to any one of many levels. When the entrepreneurs' expectations regarding the future market demand are similar, the market equilibrium price tends to converge to the level where the competitors' relative expected market shares are preserved. Thus, equilibrium market price and market shares must be explained simultaneously. To do so, we must incorporate investment into the model. Due to limitations of space, formal analysis of this must be left for another occasion; only an intuitive presentation is made here.

Let the oligopolists produce a homogeneous product and face different cost functions. Regardless of firm size, since the product is homogeneous, demand could initially be distributed randomly among all producers and each could expect an equal market share. Based on this expected demand pattern, the optimal price of a low-cost firm would be lower than that of a high-cost firm. A redistribution of demand would follow. In addition, the imputed value of the low-cost firm's capital would be higher than those of the high-cost firms. If the imputed value were to exceed the market price of capital, the low-cost firm would invest. The effect of a redistribution in demand and an increase in the capacity of the low-cost firm would cause a decrease

in market price and an increase in market share for the lowest-cost firm. Adjusting in this fashion, both the equilibrium market price and the equilibrium market share would emerge simultaneously.

2. *Market Segmentation*

It is not an uncommon phenomenon that large and small firms coexist in the same industry. The continuous presence of the small firm sector is frequently attributed to special production techniques suitable only to small scale production, the availability of small quantities of high grade raw materials, and the vigorous application of antitrust laws. While these may have been valid explanations in specific instances, my model implies that the presence of small firms may actually serve the economic interest of the large firms and that coexistence, in fact, represents a state of equilibrium.

It is true that when the market demand fluctuates mildly, the presence of the small-firm sector is detrimental to the large firms. Small firms do not accumulate inventory for the purpose of stabilizing market prices. Therefore, a decrease (increase) in market demand implies a fall (rise) in the price in the small-firm sector. Given that the large firms are committed to stabilize prices, when the market demand falls (rises) the quantity exchanged in the small-firm sector will decrease (increase) proportionately less than that in market demand. The presence of the small-firm sector thus causes the large firms to face greater variations in demand, to bear greater burdens of inventory, and to reduce their abilities to maintain prices. It can be construed as detrimental to the large firms.

However, the burdens turn into benefits when the market demand fluctuates widely. Consider the case first examined by John Maurice Clark where small firms employ production techniques which are not capital intensive and have limited economies of scale. A small initial capital investment will quickly bring long-run average cost to its minimum; a further increase in investment will sharply reduce efficiency and increase average production cost. Thus the long-run average cost curve of the small firm is V shaped. Under these conditions, an increase in market demand will induce primarily a price rise; entry into the small-firm sector is especially encouraged. Conversely, a decrease in market demand results primarily in price reduction; when price falls below the firm's short-run average variable cost, the firm will leave the market. Thus, in the small-firm sector, mild or short-lived fluctuations in market demand are met by price adjustment, while wider or sustained fluctuations are met by an adjustment in the number of firms.

The implication of my model is evident. The industry's response to market demand fluctuation will be borne initially by the large firms through expansion and contraction of inventories and possibly of production. If the upswing or downturn persists, the burden of adjustment will shift to the small-firm sector through entry and exit of firms. The change in the number of small firms will have a stabilizing effect on the large firms. The latter thus may be willing to pay the price of a slightly reduced stability during mild fluctuations in market demand in exchange for a greater stability when market demand varies widely.

D. *General Equilibrium and Resource Allocation*

Traditionally, the evil of monopoly is attributed to the fact that the monopolist sets his price above his marginal cost. This pricing policy leads to a price higher and an output lower than the competitive mechanism would determine. Resources are wasted because they are forced into alternatives which are less socially desirable. Moreover, monopoly price is often thought to be arbitrary and nonresponsive to changes in market conditions; thus, the presence of monopoly also frustrates resource allocation. When we move away from the static certainty situations, my analysis suggests that the shortcomings of monopoly may not be compared unfavorably with those of pure competition.

1. *Monopoly Output*

There are two contributing factors that cause the quantity exchanged in a market to fall short of that prescribed by the neoclassical competitive equilibrium. The first is the presence of risk-averse agents, and the second, the presence of monopoly elements in the market.

Under the condition of uncertainty, a risk-averse entrepreneur in a competitive market will produce less than that he would under certainty. (For a static analysis, see Agnar Sandmo.) Thus, under uncertainty, the market supply curve will shift to the left of that under certainty, causing an increase in the expected market price and a decrease in the quantity exchanged.

The monopolist restricts output in a different way; he balances marginal revenue instead of the higher average revenue with marginal cost. The presence of uncertainty will cause him to restrict output further under a flexible price policy regime. (For a static analysis, see Leland.) However, the monopolist, who has some influence over his uncertain environment, does have the opportunity to choose various strategies to reduce the degree of risk he faces. One of these strategies is a stable price policy. As was shown in Section I, facing a stable price, the buyer's price uncertainty is alleviated; a risk-averse buyer is thereby encouraged to purchase more of the seller's product. New buyers also may be attracted. Conjectural variation in demand among sellers is also reduced contributing to the incentive for the seller to increase production. With an increase in demand and a greater willingness to supply, the quantity produced and exchanged will necessarily increase as compared with that in the flexible price case. In some cases, the reduction in output contributed by uncertainty in a monopoly market may be negligible. Thus, we see that the monopolist restricts output primarily through balancing marginal revenue with marginal cost; while, in contrast, the competitive market restricts output through risk averseness of the entrepreneurs. Comparing the extent of output reductions in these cases, it is not at all clear which one is greater. I therefore suggest that under the condition of uncertainty, it is possible that the monopolist may not restrict output beyond the extent of the competitive market.

2. *Stable Price and Resource Allocation*

From the discussions in Section I, we may draw two relevant conclusions: (i) If a stable price is selected in order to maximize the entrepreneur's expected utility of profit over the entire decision horizon, then the selected price is not arbitrary. (ii) The selected price is only invariant as long as the entrepreneur's expectation about the market condition does not change. On the basis of these observations, I suggest that while stable prices are found in conjunction with monopoly power, stable price itself does not necessarily lead to inefficient allocation of resources.

There are two ways that resources can be reallocated following a change in market conditions: market adjustment through price and sales; and internal adjustment through changes in the firm's inventory and rate of production. In a world of static certainty when the market is purely competitive, through price (and entry and exit of firms) the market bears the burden of adjustment to demand and supply variations. The marginal social cost of the product coincides with the marginal cost of production. Resource allocation is optimal when the price of the product (marginal social gain) is equal to the marginal cost of production (marginal social cost). Price flexibility is essential for optimal resource allocation.

In a world of dynamic uncertainty, the firm performs intertemporal arbitrage; the market variation is met only in part by market adjustment in sales, and the remainder is met by the firm's internal adjustment in inventory and, to a lesser degree, in production. Social marginal cost diverges from the firm's marginal cost of production. Under these circumstances, the static certainty-oriented normative criterion—price equal to marginal cost—ceases to be valid; price

flexibility is no longer necessary for optimal resource allocation.

A new welfare criterion is needed for judging optimal resource allocation. In the absence of an appropriate normative criterion, we cannot say "when a given price has become too high or too rigid to cause a misallocation of resources," or "whether market price adjustment is always superior to intrafirm inventory and production adjustment." This is especially true, as in an atomistic market where entry and exit of firms frequently accompany price adjustments. When resources are not perfectly mobile, the exit of firms in response to market decline involves abandonment of capital and, as shown by Clark, is clearly wasteful

III. Conclusion

I have argued that stable price emerges as a result of the entrepreneur's attempt to manage endogenously generated uncertainties. The major focus in Section I was the analysis of price formation and the discussion of the structure of the firm's decision problem. A detailed examination of the equilibrium conditions and comparative statics of the stable price model is deferred to elsewhere. Only some broader implications of stable prices were presented in Section II.

From the methodological point of view, it could also be concluded that the proposed stable price model serves well as a synthesizer for the theories summarized in the introduction. The model by its nature describes an entrepreneur's long-run decision problem and reflects his concern over conjectural variation in demand among rivals. It also includes the consideration that price changes are costly. Contrary to the direct cost used by Mason, in the context of this paper, cost is in terms of the opportunity profit foregone by following a flexible price policy. Finally, the proposed model also helps to reconcile the stable price phenomenon with the maximization objective of the entrepreneur. While it is true that a stable price is inconsistent with profit maximization under the neoclassical framework, by incorporating hypothesis testing into the decision process, stable price and expected utility maximization are made compatible.

This paper has been concerned primarily with positive economic analysis relating to the stable price phenomenon and has yielded few normative implications. Moreover, the meager normative implications that remain are essentially negative in nature: the neoclassical welfare criteria are no longer adequate to guide policies in an economy with uncertainty. Unlike the world of a neoclassical economy, a monopolist need not restrict output beyond what the competitive market would do. Moreover, when the entrepreneur faces an uncertain market environment and a multiperiod decision horizon, price flexibility ceases to be a necessary condition for efficient allocation of resources. In this market environment, resources are allocated via both market price adjustment and the firm's internal adjustments in production and inventory. What constitutes an optimal division of the burden of adjustment, I cannot say. Unless new welfare criteria are devised to show when a price becomes too high and to distinguish price stability from price rigidity, my analyses will provide little guidance to public policy. There is much to be done in order to make these analyses policy relevant. This paper represents only an initial step.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk Bearing*, Chicago 1971.

M. Balch and S. Wu, "Some Introductory Remarks on Behavior Under Uncertainty," in Michael S. Balch et al., eds., *Essays on Economic Behavior Under Uncertainty*, Amsterdam 1974, 1-22.

D. P. Baron, "Demand Uncertainty in Imperfect Competition," *Int. Econ. Rev.*, June 1971, *12*, 196-208.

E. K. Choi and S. Y. Wu, "Further Notes on Risk Aversion with Many Commodities," mimeo., Univ. Iowa 1977.

J. M. Clark, "Toward a Concept of Work-

able Competition," *Amer. Econ. Rev.*, June 1940, *30*, 241–56.

Richard M. Cyert and James G. March, *A Behavior Theory of the Firm,* Englewood Cliffs 1963.

William J. Fellner, *Competition Among the Few*, New York 1949.

J. Hadar and W. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.*, Mar. 1969, *59*, 25–34.

Roy F. Harrod, *Economic Essays*, New York 1952.

G. A. Hay, "Price, Production and Inventory Theory," *Amer. Econ. Rev.*, Sept. 1970, *60*, 531–45.

R. B. Heflebower, "Toward a Theory of Industrial Markets and Prices," *Amer. Econ. Rev. Proc.*, May 1954, *44*, 121–39.

L. Hurwicz, R. Radner, and S. Reiter, "A Stochastic Decentralized Resource Allocation Process, Part I," *Econometrica*, May 1975, *43*, 187–222.

M. Kurz, "The Kesten-Stigum Model and the Treatment of Uncertainty in Equilibrium Theory," in Michael S. Balch et al., eds., *Essays on Economic Behavior Under Uncertainty*, Amsterdam 1974, 389–97.

R. Lanzillotti, "Pricing Objectives in Large Companies," *Amer. Econ. Rev.*, Dec. 1958, *48*, 921–40.

H. E. Leland, "The Theory of the Firm Facing Uncertain Demand," *Amer. Econ Rev.*, June 1972, *62*, 278–91.

E. Mason, "Price and Production Policies of Large-Scale Enterprise," *Amer. Econ. Rev.*, Mar. 1939, *29*, 61–74.

Gardiner C. Means, "The Administered-Price Thesis Reconfirmed," *Amer. Econ. Rev.*, June 1972, *62*, 292–306.

———, *The Corporate Revolution in America*, New York 1962.

Joseph Orlicky, *Material Requirement Planning*, New York 1974.

R. Radner, "Existence of Equilibrium of Plans, Prices, and Price Expectations in a Sequence of Markets," *Econometrica*, Mar. 1972, *40*, 289–304.

M. Rothschild and J. Stiglitz, "Increasing Risk: I. A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43.

A. Sandmo, "On the Theory of Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

F. M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

Martin Shubik, *Strategy and Market Structure*, New York 1959.

H. A. Simon, "Theories of Decision Making in Economics and Behavioral Sciences," *Amer. Econ. Rev.*, June 1959, *49*, 253–83.

A. M. Spence, "Stable Prices, Fluctuating Demand, and Inventory Adjustments in Oligopoly Industries," mimeo., Harvard Instit. Econ. Res., Harvard Univ. 1976.

George J. Stigler and James K. Kindahl, *The Behavior of Industrial Prices*, New York 1970.

P. Sweezy, "Demand Under Conditions of Oligopoly," *J. Polit. Econ.*, Aug. 1939, *47*, 568–73.

J. F. Weston, "Pricing Behavior of Large Firms," *Western Econ. J.*, Mar. 1972, *10*, 1–18.

R. Wilson, "Informational Economies of Scale," *Bell J. Econ.*, Spring 1975, *6*, 184–95.

S. Y. Wu, "Indebtedness and Consumer Allocation Over Time," *Metroecon.*, Jan. 1974, *26*, 97–108.

———, "Price, Production, Sales, and Inventory in a Monopoly Market Under Uncertainty," mimeo., Univ. Iowa 1975.

Labor Supply Functions in a Poor Agrarian Economy

By PRANAB K. BARDHAN*

The analytical literature on employment, unemployment, and wage determination in poor agrarian economies is large, albeit inconclusive. Empirical work in this area is comparatively scanty. For the most part it relates either to the question of "surplus" labor in peasant agriculture (and other unorganized activities) or to that of labor use and productivity in studies of production functions fitted to farm management data. There have been few systematic empirical studies of labor supply and labor market participation behavior of peasant households. The usual farm management data are not good enough for this purpose, particularly because they exclude the substantial class of landless laborers who do not have a farm. In this paper I have used detailed data collected from nearly 4,900 rural households (including landless laborers, farmers, and nonagricultural workers) in West Bengal in what may be among the first econometric attempts to estimate labor supply functions[1] in peasant agriculture. The data set is part of a very large-scale employment and unemployment survey of households carried out by the National Sample Survey Organization in India for the one-year period of October 1972–September 1973. In Section I the nature of the data is described and the results presented on labor supply behavior. My evidence seems to be against the standard horizontal supply curve of labor assumed in a large part of the development literature. In Section II the factors influencing labor participation rates for rural women are analyzed. Section III contains an analysis of the wage rates quoted as "acceptable" by different groups of respondents. Such answers came in response to hypothetical questions on wage employment to give us some idea of the supply prices of labor.

I

West Bengal is a densely populated (second highest density among major Indian states), poor (44 percent of rural population having less than Rs.30—about $4—per capita per month expenditure in 1972–73 at current prices) region with paddy-jute agriculture under heavy rainfall and little irrigation. There are pockets of advanced industrialization and commercialization in the megalopolis of Calcutta and its hinterland.

The sampling design of the survey for rural West Bengal was stratified two stage,

*Professor of economics at University of California-Berkeley. I am grateful to George Borts, Bent Hansen, Lyn Squire, T. N Srinivasan, S. D Tendulkar, and a referee for comments on an earlier draft; to Nikhilesh Bhattacharyya and Sudhir Bhattacharyya for help and guidance at the stage of data collection; and to Dan Kirshner and Karen Olsen for research assistance. The Indian Council of Social Science Research financed the data collection and collation and the Ford Foundation financed research analysis and computer costs. All errors, needless to say, are mine alone.

[1]Pan Yotopoulos and Lawrence Lau have estimated labor supply functions on the basis of Indian farm management data as part of the micro foundations for their general equilibrium model of the agricultural sector. There are a number of reasons why such data are not precisely suitable for their purposes. First, farm management data exclude the landless households who contribute a substantial part of the agricultural labor supply. Second, it seems they have run regressions over cross sections containing the group average data presented in Government of India publications of farm management surveys scattered over eight districts in different parts of India. These widely scattered agro-climatic regions are not homogeneous. Moreover the *districts* were chosen purposively for the purpose of the surveys and not on a random sample basis. Finally the data were already aggregated over size classes of farms in the published reports. This is a rather unsatisfactory basis of estimating labor supply functions. Yotopoulos and Lau take care to emphasize the merely illustrative purpose of their exercise.

with villages being the first-stage units and households at the second stage. The survey period of one year was divided into four subrounds of three months each. One-fourth of sample villages were surveyed in each subround, each sample village visited only once in the year. The sample units were staggered over four subrounds so that valid estimates could be built up for each of the subrounds separately. The subround-wise comparisons will be used to capture the seasonality effects of agriculture. In all, data were collected for approximately 8,500 rural workers belonging to nearly 4,900 sample households drawn from about 500 sample villages. Thirty-eight percent of the households have cultivation of farm (either owned by themselves or leased in) as the principal occupation (the single most important source of income) and 32 percent have agricultural wage labor (on someone else's farm) as the principal occupation. Seventeen percent of households have cultivation as the secondary occupation (the second most important source of income) and 9 percent have agricultural wage labor as the secondary occupation.

The data for each household may be classified in three groups:

Data relating to characteristics of the household as a whole and also to those of each individual member of the household, whether a worker or not.

Day-to-day time disposition particulars on *each* day of the seven days preceding the data of survey (along with wage and salary earnings reported for the week) for those who are currently in the labor force (a *current* laborer is defined as one who worked[2] for at least one hour during the reference week or sought work or was available for work) and intensity of work in any activity on each day was measured in binary codes of "full' and "half" (so that what in the judgment of the respondent was half day's work was counted as half a day worked).[3]

Response to various probing (and sometimes hypothetical) questions about the long-term work pattern, job search, and minimum acceptable wage/salary rates from members of the household belonging to the *usual* labor force (a *usual* laborer is one whose "enduring status"—i.e., the status which prevailed over, say, the last one year and which is also likely to continue in future—is that of work or looking for work or availability for work).

In estimating labor supply functions the first serious problem is that of finding an appropriate wage as an independent variable. If the daily wage rate is derived as wage earnings divided by days worked, and if the days worked also appear as the dependent variable, there is the well-known measurement error problem (if days worked are too high, the wage rate definitionally will be too low, giving rise to a negative bias) and also a "simultaneity" problem because labor supplied and the wage may be jointly determined by other variables. Added to this, there is the practical problem that those who did not work at all (in the reference week) did not have a wage to report even though the wage rate in the market may have been positive. One solution is to estimate the wage rate as determined by other independent variables for the workers who reported work at some wage rate, and from this equation estimate a predicted wage rate for everybody, including the ones who did not have an actual wage to report.[4] Let us call this wage variable $\hat{W}$.

I have also found another predicted wage variable to be quite useful. I have data for the hypothetical minimum acceptable wage rate at which many respondents report willingness to do outside work (within the village). This hypothetical minimum ac-

[2]By the standard definition of work or "gainful" work, it excludes purely domestic work.

[3]Since work in agriculture cannot easily be reduced to a standard hourly pattern and since the rural respondents find it rather difficult to report hourly disposition of their time, intensity of work by full or half day has been adopted as a better measure.

[4]Robert Hall uses this procedure in his estimates of labor supply in the United States. But as Reuben Gronau and James Heckman have pointed out, under certain circumstances this procedure may lead to a bias in the estimation of labor supply parameters.

ceptable wage rate has been estimated as a proportion of the actual wage rate, the proportion being determined by other independent variables for those workers who report both wage rates. From this equation a predictable wage ratio is estimated for everybody. Call this predicted wage ratio variable $\hat{\omega}$, and let it represent the departure of the current wage rate from the minimum acceptable wage rate.

In some cases, even another alternative wage variable has been used. On the basis of the data on wage earnings and days of wage employment for each worker in a village, I have worked out an aggregate[5] weighted-average wage rate for agricultural work for the village as a whole. Call this village wage variable *VW*. Presumably, to the individual worker the *VW* variable is more exogenous and less amenable to individual control than the actual individual wage rate as derived from the individual's reported wage earnings. Also, *VW* is quite often positive even though an individual may not have reported any wage work.

A fourth alternative wage variable used is a variation on *VW*. Even if one accepts *VW* as a good approximation of the prevailing market wage rate in the village, this may not represent the *expected* wage rate relevant for labor supply decisions, since in a situation of widespread unemployment and underemployment, the probability of getting a job at the market wage rate is less than unity. I have therefore calculated an aggregate unemployment rate[6] for each village (denoted as *VUR*) and, using that to indicate the probability of getting a job, computed an expected wage rate as equal to $VW \cdot (1 - VUR)$; call this expected wage rate in the village *VEW*. For my sample *VEW* is roughly 87 percent of *VW*.

Let us now turn to the dependent variable, agricultural[7] labor days supplied in the reference week. First, the days in the week when the laborer did not work but was either seeking work or reported being available for work (presumably at the going market wage rate) should obviously be added[8] to the days actually worked to constitute labor supply from the point of view of the supplier. In the rest of this section, except when otherwise stated, the labor supply variable is used in this more general sense. Second, since many laborers also have some family (or leased-in) land to cultivate, and since the wage response and other determinants of hired-out labor and self-employed labor may be different, I have adopted two alternative kinds of supply variables, one to represent the hiring-out behavior, and the other the *total* supply of labor days including both labor offered in the agricultural wage market and labor applied on family farm.

Third, since the decisions on labor supply in rural families are quite often made on the household level rather than by each individual separately, I have also computed total household labor supply variables (aggregating the labor supply of all the working members of the household). For the corresponding wage rates, the village wage variables *VW* or *VEW* are used.

Fourth, since social and economic constraints on labor supply are different for men and women, I have tried to distinguish between their labor supply functions. In particular, some separate estimates were made in order to analyze the factors influencing the female rate of participation in the labor force. As for independent variables other than the wage variable, I shall comment on them as we present the esti-

[5]Since agricultural work in different farms for workers in one village during roughly the same fortnight when the village is surveyed is not likely to be different in kind, such aggregation may be legitimate.

[6]The unemployment rate is measured as the number of days in the reference week that the respondents in a village reported seeking work or being available for work as a proportion of the total number of days in the reference week they worked, or reported seeking work or being available for work.

[7]Only 4 percent of the total days worked by those who are cultivators by usual status was spent in nonfarm work, only 5 percent of the total days hired out by casual agricultural laborers was spent in nonfarm work.

[8]It is assumed that when cultivators or agricultural laborers reported they were seeking work or were available for work in the reference week, it was agricultural work they were after. This is not necessarily so, but is likely to be so in most cases.

mated functions.

Take first the set of casual agricultural laborers who constitute 29 percent of the usual labor force. In the reference week they supplied agricultural wage labor (including days looking for or being available for it) for about 5.6 days per worker on an average. (About 19 percent of casual agricultural laborers did not have any agricultural wage work to report in the reference week.) Let us call this individual supply of days of farm wage labor H_1. The estimated linear equation for this is

$$(1)\quad H_1 = \underset{(0.1088)}{8.2882} - \underset{(0.0180)}{0.9017\hat{\omega}} - \underset{(0.2036)}{1.0050X_1} - \underset{(0.0023)}{0.0117X_2} + \underset{(0.0218)}{0.1075X_3} + \underset{(0.0779)}{0.2195DS2} + \underset{(0.0751)}{0.3791DS3} + \underset{(0.1516)}{0.3005DR1} - \underset{(0.0692)}{0.2201DR3};$$

$$R^2 = 0.455; \quad F = 349.5$$

where $\hat{\omega}$ is, as defined before, the *predicted* ratio[9] of the minimum acceptable wage (for work inside village) to the actual current wage rate for casual farm labor; X_1 is per capita land cultivated in acres by the household; X_2 is age (in years) of the individual; X_3 is the number of dependents per earner in the household; $DS2$ and $DS3$ are dummy variables representing subrounds 2 (January–March quarter) and 3 (April–June quarter), respectively; $DR1$ and $DR3$ are dummy variables representing region 1 (the Himalayan region of West Bengal) and region 3 (the Central Plains region), respectively.[10]

Equation (1) shows that the supply of labor for hiring out is larger the higher is the current wage rate relative to the individual's minimum acceptable wage rate, and the lower[11] is the per capita land cultivated by the household, as expected. Age tends to depress hiring out and a larger number of dependents per earner seems to necessitate more hiring out. January–March and April–June quarters are relatively slack seasons in West Bengal agriculture and it seems that laborers are willing to hire out more in these seasons. All the variables in the equation are significant at less than 5 percent level and the F-value for the estimated equation is extremely high.

Let us now analyze labor supply decisions at the household level and also include some of the small cultivators along with landless laborers. For this purpose we extract from the whole sample the set of households with below 2.5 acres of cultivated land per household and with agriculture as the single most important source of household income. Clearly, these households are the major suppliers of labor (particularly wage labor) in agriculture. They constitute about half of the total rural households. Let us use H_2 to denote the aggregate supply of farm wage labor for all adults[12] per such a household in the reference week; on an average H_2 was 6.3 days

[9]The independent variables taken for the estimated equation of ω are W, X_2, X_3, X_8, X_{13}, $D2$, and $D6$; the R^2 for this equation = 0.207 and F = 52.4. These independent variables are identified later in the text.

[10]The four subrounds in which the year 1972–73 was divided are: October–December (subround 1), January–March (subround 2), April–June (subround 3), and July–September (subround 4). Of these subrounds, 1 and 4 are the relatively busy seasons, with the latter in most areas covering periods of sowing and transplanting of the main crop of winter paddy, harvesting of autumn paddy and jute, and the former covering periods of harvesting autumn and winter paddy. Subrounds 2 and 3 are relatively slack seasons when some of the minor crops (including summer paddy) are sown or harvested. The whole state of West Bengal has also been subdivided into four regions: region 1 is the Himalayan region (covering districts of Darjeeling, Jalpaiguri, and Cooch Bihar), region 2 is the Eastern Plains (districts of West Dinajpur, Malda, Murshidabad, Nadia, and Birbhum), region 3 is the Central Plains (districts of twenty-four Parganas, Howrah, Hooghly, and Burdwan), and region 4 is the Western Plains (districts of Bankura, Purulia, and Midnapur).

[11]A larger size of farm (per capita) reduces supply of labor in two ways, partly through the income effect of larger nonwage income and partly through status effect of making one more reluctant to work for others.

[12]Adults are defined as those belonging to the 15–60 age group.

in the reference week. The estimated equation for H_2 is

$$(2)\quad H_2 = \underset{(0.3852)}{0.5993} + \underset{(0.2619)}{1.6071VW} - \underset{(0.0563)}{0.3411(VW)^2} + \underset{(0.0400)}{0.2104X_3} + \underset{(0.0813)}{3.7293X_4} - \underset{(0.0988)}{3.3574X_5} - \underset{(0.1830)}{0.4438DS2} - \underset{(0.3099)}{0.6292DR1} - \underset{(0.1923)}{0.4111DR2};$$

$$R^2 = 0.508; \quad F = 406.4$$

where VW is, as defined before, the weighted-average farm wage rate in the village in the current week (for the household as a whole the general village wage rate seems more appropriate than any individual wage rate received); X_3 is, as before, the number of dependents per earner in the household; X_4 is the number of adults in the household who are in the farm labor force by *usual* status; X_5 is land cultivated by the household; $DS2$ is the dummy for subround 2 (January–March); $DR1$ and $DR2$ are dummy variables representing region 1 (the Himalayan region) and region 2 (the Eastern Plains region), respectively.

Equation (2) shows that H_2 responds positively to VW, but the coefficient for the quadratic term $(VW)^2$ is negative. As in equation (1), hiring out seems to go down with land cultivated and up with the number of dependents per earner in the household. The larger is the number of working adults in the households, obviously, more labor days are available for hiring out. Unlike equation (1), there is *less* willingness to hire out in the relatively slack season (the income effect possibly outweighted by the general job search discouragement effect) of January–March and in the Himalayan region of West Bengal. All the variables in equation (2) are significant at less than the 5 percent level and the F-value for the estimated equation is extremely high.

In equation (2) H_2 refers to the supply of farm wage labor for the adults in the household as a whole. Let us see how the estimates differ for the male adult members in the household. I define H_{2m} as the supply of farm wage labor for male adults taken together for each household belonging to the set of agricultural worker households cultivating less than 2.5 acres of land; on an average H_{2m} was 5.1 days in the reference week. The estimated equation for H_{2m} is

$$(3)\quad H_{2m} = \underset{(0.2535)}{0.7038} + \underset{(0.1843)}{1.1826VWM} - \underset{(0.0403)}{0.2317(VWM)^2} - \underset{(0.0780)}{2.7754X_5} + \underset{(0.0809)}{4.2763X_6} - \underset{(0.1525)}{1.9208X_7} - \underset{(0.1425)}{0.2362DS2};$$

$$R^2 = 0.559; \quad F = 666.2$$

where VWM is the aggregate farm wage rate in the village for male laborers, X_6 is the number of adult males in the household and X_7 is the number of adult males in the household having higher than primary education. The response of the wage rate, the quadratic term, of land cultivated, and of the dummy for subround 2 are of the same sign as in equation (2). The larger the number of adult males in the household, obviously, more labor days are available for hiring out. It is interesting to note that education discourages hiring out on agricultural wage labor: the larger is X_7, the lower is H_{2m}. (X_3, the number of dependents per earner in the household and all the region dummies turn out to be insignificant and their F-value is so low that their omission in equation (3) has improved the value of $(\bar{R})^2$.) All the variables in equation (3) are significant at less than 1 percent level except $DS2$, which is significant at 10 percent level, and the F-value for the estimated equation is extremely high.

I ran equations similar to (3) for the separate categories of landless wage labor and small (up to 2.5 acre) cultivator households. The results are very similar to equation (3), except that VWM is positive and significant for males in wage labor households, positive but statistically not very

significant for small cultivator households, and that the latter, for obvious reasons, hire out *less* in the busy quarter of October-December.

Let us now see how the results change if, in equation (3), instead of VWM, we take $VEWM$, i.e., the expected village wage rate for male agricultural laborers. This variable is calculated as VWM in the current week multiplied by the difference of the estimated village unemployment rate in the current week from unity. The estimated equation is

$$(4)\quad H_{2m} = \underset{(0.2490)}{0.8925} + \underset{(0.2024)}{1.2979VEWM} - \underset{(0.0474)}{0.3007(VEWM)^2} - \underset{(0.7789)}{2.7834X_5} + \underset{(0.8095)}{4.2714X_6} - \underset{(0.1523)}{1.9067X_7} - \underset{(0.1432)}{0.2838DS2};$$

$$R^2 = 0.559; \quad F = 666.3$$

This equation is very similar to equation (3). Similarly, in equation (2), VW may be replaced by VEW without much change.

In equations (1)-(4) I have analyzed the hiring-out behavior of agricultural laborers and small cultivators. How about the hiring-out behavior of all cultivators taken together? Let us therefore take the set of all households cultivating more than 0.1 acre of land and having cultivation as the principal occupation for the household. Let us use H_{3m} to denote the aggregate supply of farm wage labor for all male adults per such a household; on an average H_{3m} was only 1.6 days in the reference week. The estimated equation for H_{3m} is

$$(5)\quad H_{3m} = \underset{(0.3346)}{1.3809} - \underset{(0.0991)}{0.4327VWM} - \underset{(0.0297)}{0.2641X_3} + \underset{(0.0712)}{1.5667X_6} - \underset{(0.1584)}{0.4256DR2} + \underset{(0.1573)}{0.3308DS3};$$

$$R^2 = 0.180; \quad F = 101.7$$

Not only does (5) explain a much lower proportion of the variation in hiring out than, say, in (3), the wage response is now significantly negative. This is in contrast with all the earlier equations relating to agricultural laborers and small cultivators. The $DS3$ refers to the dummy for subround 3 (April-June), which is a relatively slack season, when presumably the cultivator households have less work on their own farm and are more willing to hire out labor, contrary to the case in equation (2) for less landed households. As in equation (2), there is less hiring out in region 2 (the Eastern Plains region). As expected, the supply of hired-out male labor is positively associated with the number of adult males in the household; it is not, however, clear why it is negatively associated with X_3 (it is possible that males burdened with a larger number of dependents in primarily cultivator households work more intensively on their own farm and have less time for hiring out). The coefficients of X_5, land cultivated, and X_7, the number of adult males having higher than primary education are appropriately negative, but they turn out to be insignificant and their F-value is so low that their omission in equation (5) improves the value of $(\bar{R})^2$.

For about 77 percent of this set of cultivating households, H_{3m} is zero. I therefore also tried a *logit* model of predicting the binary choice behavior of these households in terms of H_{3m} taking zero and positive values. Without going into the details let me only note that, consistent with the regression equation (5), the logit maximum likelihood estimates suggest that a smaller number of adult males in the household, a higher village average wage rate, a larger size of land cultivated, and a larger number of dependents all increase the likelihood of H_{3m} taking the value zero (the likelihood ratio is .309).

In all the equations so far I have looked at the supply of farm wage labor in the market. For the set of cultivator households taken for equation (5), the predominant component of labor supply is that of work on the family farm. Let S be the total supply of adult farm labor days (including work on family farm, hiring out, and seeking or being available for work) for the cultivating household as a whole; on an average S was 11.2 days in the reference

week. The estimated equation for S is

$$(6)\quad S = \underset{(0.2094)}{0.1882} + \underset{(0.0342)}{0.1883X_3} + \underset{(0.0793)}{5.3620X_4} + \underset{(0.0312)}{0.1522X_5} - \underset{(0.1849)}{0.8743DS2};$$

$$R^2 = 0.701; \quad F = 1354.3$$

The wage rate variable does not appear in equation (6) since it turned out to be extremely insignificant. (It is possible that for cultivating households the relevant wage variable is not the wage rate in the reference week, but the *annual* wage income as the opportunity cost of working on their own farm—for which I do not have data). The number of working adults as well as the size of the household farm is obviously positively associated with total labor supply and so is the number of dependents per earner necessitating more work. The $DS2$ referring to the dummy for subround 2 (January–March) implies smaller total labor supply in a slack season. All the independent variables are significant at less than 1 percent level and the F-value for the equation is extremely high.

In equation (6), S refers to the total supply of adult farm labor days for the cultivating household. Let us now consider the estimate for only the adult male farm labor days. Let S_m be the total supply of adult male farm labor days for the cultivating household; on an average S_m was 9.1 days in the reference week. The following *log*-linear equation for S_m gives a better fit than the linear equation:

$$(7)\quad \log S_m = \underset{(0.1327)}{-0.6164} + \underset{(0.1795)}{0.0884 \log VWM} + \underset{(0.0235)}{0.0406 \log X_3} + \underset{(0.0704)}{0.2156 \log X_5} + \underset{(0.0204)}{1.1349 \log X_6} - \underset{(0.0104)}{0.0955 \log X_7} + \underset{(0.1085)}{0.3140DS1} + \underset{(0.0949)}{0.3304D1};$$

$$R^2 = 0.615; \quad F = 615.6$$

where the variables are as defined before except $D1$ which is a dummy to indicate if the village is predominantly multiple crop producing. Note that VWM has a positive but statistically insignificant response. As expected, S_m is positively associated with the number of dependents per earner, the number of adult males in the household, the size of the family farm, the dummy for the village multiple cropping and the dummy for subround 1 (October–December) which, as noted before, is a relatively busy season. As before, education seems to discourage agricultural work: S_m is negatively associated with the number of adult males in the household having more than primary education. Except for VWM, the coefficients for all the other variables in equation (7) are statistically significant (at less than 5 percent level for X_5, X_6, X_7, $DS1$, and $D1$, and at less than 10 percent level for X_3) and the F-value for the equation is extremely high.

II

Let us now focus on the participation rates for women. My equations estimate only a small part of the variations here, but they are at least able to identify some of the important determinants. Let us first take the set of households whose primary source of income is from cultivation or agricultural labor. Define P_1 as the total number of days in all kinds of "gainful" work (except household chores) in the reference week per adult woman in such a household; on an average P_1 has the very low value of 0.13 days (with a standard deviation of 0.29). The estimated equation for P_1 is

$$(8)\quad P_1 = \underset{(0.0289)}{0.1548} - \underset{(0.0064)}{0.0383VW} - \underset{(0.0021)}{0.0317X_3} - \underset{(0.0052)}{0.0086X_6} - \underset{(0.0002)}{0.0007X_8} + \underset{(0.0003)}{0.0021X_9} + \underset{(0.0109)}{0.0297DS1} + \underset{(0.0114)}{0.0786DS4} - \underset{(0.0092)}{0.1215D2} - \underset{(0.0196)}{0.1053VUR};$$

$$R^2 = 0.164; \quad F = 90.0$$

where the new independent variables are

X_8, per capita expenditure of the household; X_9, a general agricultural development (composite) index for the district where the household is located;[13] $D2$ is a *dummy variable to separate the lowest group in the social hierarchy (scheduled castes and scheduled tribes) from the rest of the population, and VUR is the village unemployment rate.*

An important factor in explaining variation in P_1 seems to be D2: low-caste and tribal women participate more than higher caste women. The wage response is significantly negative.[14] As expected, P_1 is negatively associated with the number of dependents (necessitating more household work), with the number of adult males in the household (the smaller their number the greater is the necessity for women to join the labor force), with the village unemployment rate (discouraging work participation), and with the family standard of living. Women seem to participate more in agriculturally developed areas—note the significant positive association between P_1 and X_9—possibly due to increased work opportunities and the liberating effects of attendant commercialization. The variable P_1 is positively associated with the dummies for subround 1 (October–December) and subround 4 (July–September), possibly because these include the relatively busy seasons of transplanting, harvesting, and threshing paddy (transplantation and threshing are quite often an exclusively female job in this area). Except for X_6 all of the variables in the equation are significant at less than 1 percent level. (I have also tried a *logit* model of predicting the binary choice behavior of cultivator households in terms of P_1 taking zero and positive values. The same variables which are important in the regression equation (8) are again significant in the same way in the logit maximum likelihood estimates—the likelihood ratio was 0.469.)

In equation (8) I looked at the participation rates for the women in the household as a whole. Now I shall analyze the participatory behavior of individual women and also leave out the large number of women whose usual activity is only doing household chores. If attention is thus confined to adult women who are workers by *usual* status, the number of women in this set is 24 percent of the total number of adult women. Suppose P_2 is number of days in the reference week each woman in this set either worked, sought work, or was available for work; on an average P_2 was five days in the reference week. The estimated equation for P_2 is as follows:

$$(9)\quad \begin{aligned} P_2 = & -6.2753 + 9.6440\hat{W} \\ & \quad (2.5412) \quad (2.4050) \\ & - 2.0851(\hat{W})^2 - 0.3006D2 \\ & \quad (0.5670) \qquad (0.1468) \\ & + 0.4549DS4 + 0.4742D3 \\ & \quad (0.2192) \qquad (0.1478) \\ & + 0.5186D4; \\ & \quad (0.1607) \\ & R^2 = 0.049; \qquad F = 9.8 \end{aligned}$$

where $\hat{W}$ is the predicted farm labor wage rate,[15] $D3$ is the dummy for usual nonfarm workers (as contrasted with usual farm workers), and $D4$ is the dummy to represent those women who are widowed, divorced, or separated (as contrasted with other marital status). The variable P_2 seems

[13]The Reserve Bank of India *Bulletin* gave a composite index of agricultural development for most districts in India around the middle of the 1960's. This index is a weighted-average index of a number of agroclimatic and commercialization indices related to (a) the percentage of gross cropped area irrigated, (b) multiple cropping intensity, (c) average rainfall, (d) soil rating, (e) percentage of area under cash crops, and (f) surplus of cereals over consumption in the district. I have used the composite indices given there for the different districts in West Bengal, 1964–65, to distinguish households in the sample located in areas with different agro-climatic environments.

[14]I should also note that the wage variable VW is a weighted-average village wage for males and females taken together. Since the males supply most of the wage labor, it reflects more the male wage; I am assuming that the male to female wage ratio is similar across villages.

[15]The independent variables which are significant in the estimated equation for W are X_1, X_2, X_3, X_4, $D6$, $D7$, $DS1$, $DS2$, and $DS4$; the R^2 for this equation = 0.106 and F = 35.1.

to respond positively to $\hat{W}$, but at a declining rate. It is interesting to note that women who are widowed, divorced, or separated, participate more in labor supply, and women belonging to higher castes participate less. Nonfarm workers seem to have a higher participation. Subround 4 (July–September) involving the busy season of harvesting (of the autumn crop) and transplanting brings forth more women participation. These independent variables, while they are all significant at less than 5 percent level, explain, however, only a small part of the variation in P_2.

To briefly summarize the results of the last two sections, the wage response of labor supply seems to be significantly positive for the set of agricultural laborers and small cultivators, and also for that of women in the *usual* labor force. The wage response is not significant for *total* labor supply for the set of cultivators of all size groups taken together. There seems to be some evidence for at least a locally backward-bending supply curve of labor for the set of all adult women (primarily housewives) and of hired-out farm labor for the set of cultivators of all size groups taken together.

Even when the wage response of hiring out farm labor is positive, as in the case of agricultural laborers and small cultivators, it is not very large. According to my estimates, the elasticity of H_2 with respect to VW is 0.29 and that of H_2 with respect to VEW is about 0.2. These elasticity values are drastically different from the infinite elasticity presumed in the horizontal supply curve of farm labor used in a large part of the development literature. It is possible, however, that the presumed horizontal supply curve of labor in the literature is not an *ex ante* supply curve as in my estimates, but more like a locus of equilibrium points, as in the case of the Keynesian aggregate supply curve, suggesting constancy of the equilibrium wage rate. But, as I have shown in my 1977 paper, on the basis of the same data set for rural West Bengal, there is a considerable degree of dispersion in the agricultural wage rates, and the variations are explainable largely in terms of demand and productivity factors, contrary to the implications of the hypothesis of wage constancy.

In general, it seems labor supply is primarily determined by other economic, social, and demographic constraints, and is not highly responsive to the wage rate. Some, though certainly not all, of these constraints are reflected in some of the independent variables other than wage in the regression equations. Hiring-out behavior is usually related positively to the number of adult workers in the family and the number of dependents per earner (except in the case of all cultivators); it is negatively related to the size of land cultivated by the household, the level of living, or the educational level of adults in the household. Total labor supply (including supply on family farm) of cultivators also behaves similarly, except that now, understandably, it is positively associated with the size of land cultivated by the household, village multiple cropping index, and with the busy season and negatively with the slack season. Labor participation of women is positively associated with lower caste families, the district agricultural development index, widowed (or divorced or separated) women, and with the busy season. It is negatively associated with number of dependents, with number of adult males in the household, the village unemployment rate, and the family standard of living.

III

Apart from reporting wage and salary earnings actually received in the reference week, the respondents in the data set also answered hypothetical questions about an acceptable wage rate (in case they were looking for wage employment) inside and outside the village. It would be interesting to analyze this information on what might be considered as *ex ante* supply prices on the part of workers. Take the set of casual agricultural laborers. About 30 percent of them say that they are willing to accept wage employment inside or outside the vil-

lage (about 22 percent are willing to accept wage employment inside but not outside the village). For these laborers who reported both a current farm wage received and an acceptable wage for a hypothetical wage employment inside the village, the value of the ratio of the acceptable to the actual wage is estimated to be about 1.6; 60 percent thus seems to be on an average the desired margin of wage on a hypothetical job inside the village over their current one.[16] Similarly, 87 percent is estimated to be the desired margin of wage on a hypothetical job outside the village over the current one for those who reported preparedness to accept jobs outside the village.[17]

Let us call A_c the acceptable wage rate reported by casual agricultural laborers for hypothetical wage employment inside the village. For those who report it, the mean value of A_c is Rs.3.73 with a standard deviation of 1.61. Let us now try to find out the factors which may influence the level of A_c. The estimated equation for A_c that follows leaves unexplained most of the variations in A_c, but is at least able to identify some of the factors bearing on it:

$$(10)\quad A_c = \underset{(0.1464)}{3.1170} + \underset{(0.0444)}{0.1810W} + \underset{(0.0337)}{0.0687X_3} + \underset{(0.1771)}{0.3515X_{13}} - \underset{(0.1256)}{0.4401D6} + \underset{(0.1036)}{0.2176D2};$$
$$R^2 = 0.041;\quad F = 12.1$$

where the new variables are X_{13}, which is a measure[18] of the rate of underemployment for casual agricultural laborers, and $D6$, which is a dummy variable taking a value of 1 in the case of women. All the independent variables are significant at 5 percent level, and although the R^2 is very low, the F-value for the equation is significant at less than 1 percent level. Predictably, the acceptable wage rate is larger the higher is the current wage rate received, the larger is the number of dependents per earner (indicating greater need) in the household and the higher is the caste status (presumably implying more bargaining power) the individual enjoys; and it is lower in the case of women. But what is not so clear is why A_c is positively associated with the measure of current underemployment for the individual (it is possible that those who are severely underemployed for some time have already got themselves adjusted to the rhythms of various kinds of domestic work or collection activities for "free" goods from common property in the village, so that a higher than usual wage is required to bring them back to the wage labor market).

The variable A_c is the acceptable wage rate inside village for casual agricultural laborers. We have also tried to explain variations in a similar reportedly acceptable wage rate inside village for self-employed farmers and family helpers—let us call this wage rate A_s. For those who report it, the mean value of A_s is Rs.4.27 which is higher than the mean of A_c for casual agricultural laborers. The estimated equation for A_s again leaves unexplained most of the variations in it, but is again able to identify some of the factors bearing on it:

$$(11)\quad A_s = \underset{(0.2095)}{3.3395} + \underset{(0.0288)}{0.0769X_3} + \underset{(0.0677)}{0.2201VW} + \underset{(0.4572)}{1.6057VUR} - \underset{(0.2898)}{1.4306D6} + \underset{(0.4084)}{1.2624D8};$$
$$R^2 = 0.118;\quad F = 13.6$$

The new variable in equation (11) is $D8$, which is a dummy to distinguish those having more than primary education level from others. In (11) all the independent variables are significant at less than 1 percent level

[16] Forty-one percent of the males among these laborers quoted an acceptable wage of Rs.3.5 or below and the rest quoted a figure higher than Rs.3.5.

[17] In the survey neither the nature of the job (whether agricultural, factory, public works, etc.) or its location (if shifting from current residence is necessary or not, etc.) was specified.

[18] The underemployment rate has been measured as the number of days the laborer reported seeking or being available for work as percentage of the total number of days he worked *plus* the number of days he reported seeking or being available for work.

and, although the R^2 is low, the F-value is significant at less than 1 percent level.

As expected, A_s is positively associated with the number of dependents per earner in the household, with the village wage rate, with the level of education of the respondent, and negatively associated with women. Again, as in the case of equation (10), the positive response to VUR, the village unemployment rate, is not entirely clear and may have a similar explanation.

In this paper I have analyzed a very large set of cross-sectional data for agricultural workers in rural West Bengal in estimating farm labor supply functions. Little evidence was found for the standard horizontal supply curve of labor, frequently assumed in the theoretical literature on development. The wage elasticity of supply turns out to be rather small. It seems that labor supply (as well as reported supply price for extra work) is primarily determined by social and demographic conditions for the labor supplying household and its asset situation (as reflected in the size of the household farm). One major limitation, however, of the survey data that I have used is that they are from an employment survey and do not allow us to place the labor supplying household in the whole network of interlocking relationships with other participants in the village economy. For this purpose what is needed is more intensive, possibly small-scale, micro surveys of not only the employment, social, and demographic characteristics of laborers, but also an integrated picture of the various terms and conditions of contracts in labor and related markets (land, credit, and commodity markets, in particular) where the labor household participates, in the context of some general background information on the asset distribution and the structure of economic and political power in the village. For the results of a survey of the interlinkages of land, labor, and credit relations in the agriculture of Eastern India, see the author and A. Rudra.

REFERENCES

P. K. Bardhan, "Wages and Unemployment in a Poor Agrarian Economy," *J. Polit. Econ.*, forthcoming.

______ **and A. Rudra,** "On the Interlinkages of Land, Labor and Credit Relations in Agriculture," *Econ. Polit. Weekly*, Feb. 1978, *13*, annual issue, 367–84.

R. E. Hall, "Wages, Income, and Hours of Work in the *U.S.* Labor Force" in Glenn Cain and Harold Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973.

J. Heckman, "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

R. Gronau, "The Intra-family Allocation of Time: The Values of the Housewives' Time," *Amer. Econ. Rev.*, Sept. 1973, *63*, 634–51.

P. A. Yotopoulos and L. J. Lau, "On Modeling the Agricultural Sector in Developing Economies," *J. Develop. Econ.*, June 1974, *1*, 105–27.

Government of India, National Sample Survey Organization, *Employment and Unemployment Survey*, 1972–73 (27th Round).

Reserve Bank of India, *Bulletin*, Oct. 1969, *23*, 1595.

The Design of an Optimal Insurance Policy

By ARTUR RAVIV*

Almost every phase of economic behavior is affected by uncertainty. The economic system has adapted to uncertainty by developing methods that facilitate the reallocation of risk among individuals and firms. The most apparent and familiar form for shifting risks is the ordinary insurance policy. Previous insurance decision analyses can be divided into those in which the insurance policy was exogenously specified (see John Gould, Jan Mossin, and Vernon Smith), and those in which it was not (see Karl Borch, 1960, and Kenneth Arrow, 1971, 1973). In this paper, the pioneering work of Borch and Arrow—the derivation of the optimal insurance contract form from the model—is synthesized and extended.

The incentive to insure and insurance decisions have been treated extensively by Gould, Mossin, and Smith. They analyzed the problem of rational insurance purchasing from the point of view of an individual facing a specific risk, given his wealth level and preference structure. In their analysis the individual is offered an insurance policy specifying the payment to be received from the insurance company if a particular loss occurs. The individual may choose the level of the deductible, the level of the maximum limit of coverage, or the fraction of the total risk which is to be insured. Since the premium paid by the individual is directly related to the features chosen, the optimal insurance coverage involves balancing the effects of additional premium against the effects of additional coverage. In this approach the terms of the policy are assumed to be exogenously specified and are imposed on the insurance purchaser.

Borch (1960) was the first to take the more general approach of deriving the optimal insurance policy form endogenously. He sought to characterize a Pareto optimal risk-sharing arrangement in a situation where several risk averters were to bear a stochastic loss. This framework was then used by Arrow (1971) to obtain Pareto optimal policies in two distinct cases: 1) if the insurance seller is risk averse, the insured prefers a policy that involves some element of coinsurance; (i.e., the coverage will be some fraction (less than 1) of the loss); and 2) if the premium is based on the actuarial value of the policy plus a proportional loading (i.e., the insurer is risk neutral) and the insurance reimbursement is restricted to be nonnegative, the insurance policy will extend full coverage of losses above a deductible. Arrow (1973) extended this result to the case of state dependent utility functions. In this case, the optimality of a deductible which depends upon the state was proved. Robert Wilson also dealt with the endogenous determination of optimal risk-sharing arrangements, focusing on the incentive problem and the existence of surrogate functions. Consequently, constraints on the contract or costs associated with contracting were not included.

The purpose of this paper is to explain the prevalence of several different insurance contracts observable in the real world. The previous studies addressed this issue with a diversity of underlying assumptions and, therefore, the essential ingredients of the model that give rise to the insurance policy's various characteristics are not clear. For example, in Arrow's 1971 paper it is unclear whether an insurance policy with a deductible is the consequence of risk neutrality of the insurer, nonnegativity of the insurance coverage, or loading on the premium. Could a deductible be obtained when the insurer is risk averse? Could the loading be interpreted as risk premium? Could we explain the prevalence of deduct-

*Associate professor of economics, Carnegie-Mellon University and Tel-Aviv University. I would like to thank E. Green, M. Harris, R. Townsend, T. Romer, D. Epple, and G. Constantinides for several helpful discussions.

ibles and coinsurance in insurance policies? These and other questions can be answered only from a general formulation of the insurance problem, a formulation in which the previous models are imbedded.

In this paper I undertake the development of such a model. Using the same basic framework as Borch (1960), and Arrow (1971, 1973) the choice of contracts subject to various restrictions on the class of feasible contracts is considered. The analysis generalizes and extends the previous results in several directions: 1) The form of the Pareto optimal insurance contract is identified under general assumptions regarding the risk preferences of both the insurer and insured. The necessary and sufficient conditions leading to deductibles and coinsurance are investigated. 2) The cost of insurance is explicitly recognized and shown to be the driving force behind the deductible results. This clarifies the results obtained by Arrow (1971). 3) I show the conditions under which an insurance policy with an upper limit on coverage is adopted. 4) All results are extended to the case where more than one loss can occur during the period of insurance protection. A thorough understanding of the above issues not only contributes to our understanding of insurance policies, but provides a foundation for the analysis of optimal contracts in more general situations.

In this paper, the insurance policy is characterized by the premium paid by the insured and by a coverage function specifying the transfer from the insurer to the insured for each possible loss. The admissible coverage functions are restricted to be nonnegative and less than the size of the loss. Provision of the insurance is costly, with the cost consisting of fixed and variable (depending on the size of the insurance payment) components. The premium depends on the insurance policy and the insurance cost through a constraint on the insurer's expected utility of final wealth.

It is shown that the Pareto optimal insurance policy involves a deductible and coinsurance of losses above the deductible. The deductible is strictly positive if and only if the cost of insurance is a function of the insurance coverage. In other words, if the cost of providing insurance is independent of the insurance coverage then the Pareto optimal contract does not have a deductible. I conclude that the deductible clause in insurance policies exists due to two sources: the nonnegativity constraint on the transfer from the insurer to the insured and the variable insurance cost. The coinsurance arrangement is due to either the risk aversion of the insurer or the nonlinearity of the insurance costs. The exact functional relationship for the coinsurance is derived. These results are generalizations of Arrow's 1971 work and point out the crucial assumptions underlying his results. (Arrow's results are included as special cases.) Contrary to what might be inferred from Arrow it is shown that 1) insurer's risk neutrality is neither a necessary nor sufficient condition for a policy to have a deductible; 2) insurer's risk aversion is not a necessary condition for coinsurance; 3) an optimal policy may involve a deductible and a coinsurance. A policy with an upper limit on coverage, a feature common in major medical, liability, and disability insurance, is shown not to be Pareto optimal. To explain the prevalence of upper limits, a model is provided in which a risk-averse insurer is restricted (by regulation) in determining his premium by an actuarial constraint.

These results are initially derived under the assumption that at most one loss can occur during the period of insurance protection. When several losses are allowed to occur during the insurance period (from a single or several perils), the optimal insurance policy should be based on the aggregate loss and possess the same characteristics as previously discussed. If a policy with a deductible was optimal in the single loss case, then the policy should stipulate a deductible from the total claims when two or more losses occur. Similarly, in the case of insurance with upper limits, the upper limit should apply to the aggregate loss of the insured and not to each loss separately.

The assumptions and the model are specified in the next section, followed by a char-

acterization of optimal insurance coverage when the premium is exogenously fixed. Section III completes the determination of Pareto optimal insurance policies. The behavior of a risk-averse insurer is investigated in Section IV while Section V extends all previous results to the multiple loss case. Conclusions and a critique are contained in the last section.

I. Assumptions and the Model

The insurance buyer faces a risk of loss of x, where x is a random variable with probability density function $f(x)$. Assume that $f(x) > 0$ for $0 \leq x \leq T$.[1]

The insurance policy is characterized by the payment, denoted by $I(x)$, transferred from the insurer to the insured if loss x obtains. Let us refer to $I(x)$ as the *insurance policy* or as the *coverage function*. Any admissible coverage function satisfies

$$(1) \qquad 0 \leq I(x) \leq x \qquad \text{for all } x$$

This constraint reflects the assumption that an insurance reimbursement is necessarily nonnegative and cannot exceed the size of the loss. The latter implies that the insured cannot gamble on his risk. This constraint also implies $I(0) = 0$; there is no reimbursement if there is no loss. The price paid by the insured, *the premium*, is denoted by P. Provision of insurance is costly due to administrative or other expenses and this cost is a deadweight loss relative to the insurer and the insured. It is assumed that the cost consists of fixed and variable (depending on the size of the insurance payment) components; $c(I)$ denotes the *cost* when the insurance payment is I with

$$(2) \qquad c(0) = a \geq 0, \qquad c'(\cdot) \geq 0, \qquad c''(\cdot) \geq 0$$

The insurer is assumed to maximize the expected value of his utility, which is a concave function of wealth; $V(W)$ denotes the utility function of the insurer with $V'(W) > 0$ and $V''(W) \leq 0$ for all W. Thus, the insurer is assumed to be risk averse (but not necessarily strictly risk averse). The special case of risk-neutral insurer, $V''(W) = 0$, is of particular interest.

If W_0 denotes the initial wealth of the insurer, then after selling the insurance policy and receiving the premium P, his final wealth is $W_0 + P - I(x) - c(I(x))$ if the loss x obtains. In other words, the insurer exchanges his initial certain utility $V(W_0)$ for the expected utility $E\{V[W_0 + P - I(x) - c(I(x))]\}$. A necessary condition for the insurer to offer such a policy is

$$(3) \qquad E\{V[W_0 + P - I(x) - c(I(x))]\} \geq V(W_0)$$

In the special case of a risk-neutral insurer, the risk premium equals zero and equation (3) takes the form: $P \geq E[I(x) + c(I(x))]$. Here, the policy is evaluated by the insurer according to the actuarial value of the coverage and cost. Often, in the insurance literature, it is assumed that

$$(2') \qquad a = 0, \; c'(I) = l \qquad \text{for all } I$$

That is, the costs are proportional to the insurance payment (fixed percentage loading l). In this case, the constraint on the policies offered is:

$$(3') \qquad P \geq (1 + l)E[I(x)]$$

On the insurance demand side, the insured is assumed to maximize the expected value of his utility of wealth. The insured's utility function of wealth is denoted by $U(w)$ with

$$(4) \qquad U'(w) > 0, \; U''(w) < 0 \qquad \text{for all } w$$

If w is the initial level of wealth, x the loss (a random variable), $I(x)$ the payment received from the insurer when loss x occurs, and P the premium paid for the insurance coverage, then the insured's final wealth is $w - P - x + I(x)$. Without purchasing insurance, his final wealth is $w - x$ when the loss x occurs. Thus, a necessary condition for purchasing the coverage $I(x)$ for a premium P is

$$(5) \qquad E\{U[w - P - x + I(w)]\} \geq E\{U[w - x]\}$$

[1] We could assume $f(x) \geq 0$. However, this would complicate the exposition without adding any content.

Necessary conditions for a *given* insurance contract to be acceptable to each party were given above. In order for a contract to be acceptable to both sides, both (3) and (5) have to be satisfied. In what follows, we will assume that the set of acceptable insurance contracts which satisfy these necessary conditions is nonempty. From this set, a Pareto optimal insurance policy will be chosen.

To find the form of the Pareto optimal insurance contract, we find the premium P and the function $I(\cdot)$ that maximize the insured's expected utility of final wealth subject to the constraint that the insurer's expected utility is constant. The problem is stated as follows:

$$(6)\quad \max_{P,I(x)} \overline{U}(P,I) \equiv \int_0^T U[w - P - x + I(x)]\, f(x)dx$$

subject to (1) and

$$(7)\quad \overline{V}(P,I) \equiv \int_0^T V[W_0 + P - I(x) - c(I(x))]\, f(x)dx \geq k$$

where k is a constant and $k \geq V(W_0)$.

The above problem is solved in two steps. First, in Section II, the premium P is assumed fixed and the form of the optimal insurance coverage is found as a function of P. Second, in Section III, the optimal P is chosen, thus completing the solution to our problem.

II. Optimal Insurance Coverage for a Fixed Premium

The next theorem characterizes the solution to equation (6) subject to constraints (1) and (7) when P is fixed. The theorem states that optimal insurance policies have one of two possible forms: there is either a deductible provision coupled with coinsurance of losses above the deductible, or there is full coverage of losses up to a limit and coinsurance of losses above that limit. Coverage functions satisfying (8) below are referred to as *policies with a deductible*. The *deductible* $\bar{x}_1$ is the largest loss not covered by the insurance policy. Policies satisfying (9) are referred to as policies with an upper limit on full coverage or *policies with upper limit*. The *upper limit* $\bar{x}_2$ is the largest loss which is fully covered by insurance. Usually, the coinsurance level is the proportion of the loss covered by insurance. In our analysis this proportion varies with the size of the loss. Consequently, the *coinsurance* is defined as the marginal coverage, $I^{*\prime}(x)$. From (10) it is seen that the coinsurance depends on the risk preferences of the insurer and the insured as well as on the cost function $c(\cdot)$.[2]

THEOREM 1: *The solution* $I^*(x)$ *to equation* (6), *subject to constraints* (1) and (7) *when* P *is fixed, takes one of the two forms* (8) *or* (9) *where*

$$(8)\quad \begin{aligned} I^*(x) &= 0 && \text{for } x \leq \bar{x}_1 \\ 0 < I^*(x) &< x && \text{for } x > \bar{x}_1 \end{aligned}$$

$$(9)\quad \begin{aligned} I^*(x) &= x && \text{for } x \leq \bar{x}_2 \\ 0 < I^*(x) &< x && \text{for } x > \bar{x}_2 \end{aligned}$$

In both cases, in the range where $0 < I^*(x) < x$, the marginal coverage satisfies

$$(10)\quad I^{*\prime}(x) = \frac{R_U(A)}{R_U(A) + R_V(B)(1 + c') + c''/(1 + c')}$$

where

$$A = w - P - x + I^*(x)$$
$$B = W_0 + P - I^*(x) - c(I^*(x))$$

$R(\cdot)$ denotes the index of absolute risk aversion, and c', c'' are evaluated at $I^*(x)$.

PROOF:

Constraint (7) is binding at the optimum. Since the specified problem is solved via optimal control theory we rewrite (7) as

$$(11)\quad \begin{aligned} \dot{z}(x) &= V[W_0 + P - I(x) - c(I(x))]\, f(x) \\ z(0) &= 0 \\ z(T) &= k \end{aligned}$$

Using $I(x)$ as the control variable and

[2]When we have no contraints on the insurance function and when $c(\cdot) \equiv 0$ then (10) is identical to the sharing rule given by Wilson.

$z(x)$ as the state variable, the Hamiltonian for this problem is

$$H = \{U[w - P - x + I(x)] + \lambda V[W_0 + P - I(x) - c(I(x))]\} f(x)$$

Since the Hamiltonian does not depend on the state variable it is clear that the auxilary function λ is constant with respect to x.

The necessary conditions for the optimal coverage function to maximize the Hamiltonian subject to constraint (1) are

$$(12) \quad I^*(x) = 0 \quad \text{if } \mathbf{J} \equiv U'(w - P - x) - \lambda V'(W_0 + P - a)(1 + c'(0)) \leq 0$$

$$(13) \quad I^*(x) = x \quad \text{if } \mathbf{K} \equiv U'(w - P) - \lambda V'(W_0 + P - x - c(x))(1 + c'(x)) \geq 0$$

$$(14) \quad U'[w - P - x + I^*(x)] - \lambda V'[W_0 + P - I^*(x) - c(I^*(x))][1 + c'(I^*(x))] = 0 \quad \text{for } 0 < I^*(x) < x$$

First, note that either (12) or (13) has to occur for some x, although both cannot be satisfied simultaneously. This follows directly from the fact that $\mathbf{J}$ (as defined in (12)) is continuous and increasing in x while $\mathbf{K}$ is continuous and decreasing in x. If

$$\mathbf{L} \equiv U'(w - P) - \lambda V'(W_0 + P - a)(1 + c'(0)) \geq 0$$

then (12) cannot obtain for $x > 0$. If $\mathbf{L} \leq 0$ then (13) cannot obtain for $x > 0$. Hence the optimal solution satisfies either (12) and (14), or (13) and (14). Define $\bar{x}_i$, $i = 1,2$ from

$$(12') \quad U'(w - P - \bar{x}_1) - \lambda V'(W_0 + P - a)(1 + c'(0)) = 0$$

$$(13') \quad U'(w - P) - \lambda V'(W_0 + P - x_2 - c(\bar{x}_2))(1 + c'(\bar{x}_2)) = 0$$

Clearly, $\bar{x}_i$ is uniquely defined by (12′) and (13′), respectively. (The special case $\bar{x}_1 = \bar{x}_2 = 0$ occurs if $U'(w - P) - \lambda V'(W_0 + P - a)(1 + c'(0)) = 0$.) As a result, the optimal coverage function takes one of the two forms (8) or (9). In both cases, for $x > \bar{x}_i$ (14) is satisfied. Differentiating with respect to x and using the earlier definitions of A and B we obtain for $x > \bar{x}_i$:

$$(15) \quad U''(A)[I^{*\prime}(x) - 1] + \lambda V''(B)[1 + c'(I^*(x))]^2 I^{*\prime}(x) - \lambda V'(B)c''(I^*(c))I^{*\prime}(x) = 0$$

Substituting λ from (14) and solving for $I^{*\prime}(x)$, (10) is obtained.

Notice also that, since the Hamiltonian does not depend on the state variable, the sufficiency theorem concavity requirement as shown by Morton Kamien and Nancy Schwartz is satisfied trivially. Hence, $I^*(x)$ satisfies the necessary and sufficient conditions of optimality.

Theorem 1 is easily interpreted. In the absence of constraint (1), risk aversion of both parties implies that Pareto optimal coverage involves sharing the risk according to the sharing rule (10). Equation (10) is a differential equation which, together with a boundary condition, results in a coverage function. Whether the optimal policy has a deductible or an upper limit depends on the appropriate boundary condition. For example, if the boundary condition is $I(\bar{x}_1) = 0$, then the policy has a deductible. The appropriate boundary condition depends on the fixed premium. When the premium is P, let $I_P(x)$ denote the function which solves the differential equation (10) with the boundary condition $I_P(0) = 0$. Since $0 < I'_P < 1$, this function also satisfies constraint (1). To verify whether this function is the solution to the problem the insurer's expected utility $\bar{V}(P, I_P)$ must be evaluated. Three cases could occur: 1) If $\bar{V}(P, I_P) = k$, then (7) is satisfied, $I_P(0) = 0$ is the appropriate boundary condition, and I_P is the optimal coverage function. 2) If $\bar{V}(P, I_P) < k$, then I_P is not the solution since (7) is violated. To increase the insurer's expected utility to the required level, the payment to the insured has to be reduced for some losses. This could be achieved by a boundary con-

dition specifying that $I^*(0)$ is negative. However, the constraint $I^*(x) \geq 0$ becomes binding and the appropriate boundary condition is $I^*(x) = 0$ for $x \leq \bar{x}_1$. In this case, the optimality of the deductible policy is obtained. 3) If $\bar{V}(P, I_P) > k$, the coverage can be increased. This increases the insured's expected utility, while the constraint (7) on insured's utility is not violated. The appropriate boundary condition is $I^*(0) > 0$, which together with the constraint $I(x) \leq x$, results in $I^*(x) = x$ for $x \leq \bar{x}_2$. In this case, the policy with an upper limit is obtained.

Let P_0 be the fixed premium corresponding to the first case above and let $I_0(\cdot)$ be the function solving (10). Thus, $\bar{V}(P_0, I_0) = k$ and the coverage function has the property $\bar{x}_1 = \bar{x}_2 = 0$ (i.e., (P_0, I_0) is the policy with no deductible or upper limit provision). Denote

(16) $$S_1 = \{P \mid \bar{V}(P, I_P) \leq k\}$$

and $$S_2 = \{P \mid \bar{V}(P, I_P) \geq k\}$$

By definition, $P_0 \in S_i$ for $i = 1$ and 2. Lemma 1 summarizes the discussion above and states that the optimal coverage involves a nontrivial deductible if $P \in S_1$ and a nontrivial upper limit if $P \in S_2$.

LEMMA 1: *For* $P \in S_i$, $i = 1$ *or* 2 *and* $P \neq P_0$, $I^*(x)$ *is specified by* (8) *and* (10) *or* (9) *and* (10), *respectively, with* $\bar{x}_i > 0$, $i = 1$ *or* 2.

Lemma 2 determines the effect of a change in $\bar{x}_i$ on $I^*(x)$ for $P \in S_i$. It is stated that for policies with a deductible, the coverage function decreases with the deductible level. Similarly, for policies with upper limit, the coverage function increases with the upper limit. If the sharing proportion $I^{*\prime}$ was constant, then these results clearly follow from the changes in the initial conditions of the differential equation. The proof that it is correct in the present, more general, case is given in Appendix A.

LEMMA 2: *a) If* $P \in S_1$ *then* $\partial I^*/\partial \bar{x}_1 < 0$ *for* $x > \bar{x}_1$ *and b) If* $P \in S_2$ *then* $\partial I^*/\partial \bar{x}_2 > 0$ *for* $x > \bar{x}_2$.

III. Pareto Optimal Insurance Policy

In the previous section the premium was assumed fixed. Therefore, in Theorem 1, $\bar{x}_1$, $\bar{x}_2$ and $I^*(x)$ are functions of P. We proceed to determine P^*, thus completing the determination of the Pareto optimal insurance policy. Theorem 2 proves that the search for the optimal premium can be restricted to the subset S_1 of premiums which generate policies with a deductible. Within this subset, Theorem 3 characterizes the necessary and sufficient conditions for the deductible to be nontrivial. These two results together complete the derivation of Pareto optimal policies and allow us to clearly distinguish the cases under which we would expect to observe deductibles and coinsurance clauses in insurance contracts. The results obtained by Arrow (1971) are treated as special cases thus allowing us to focus on the specific assumptions which generate these results.

The next theorem states that any insurance policy with an upper limit is dominated by the policy (P_0, I_0) with zero upper limit. In other words, the pure sharing arrangment dominates any policy with an upper limit. Intuitively, starting with the (P_0, I_0) policy, any increase in $\bar{x}_2$ (from $\bar{x}_2 = 0$) has the effect of increasing insurance coverage for all losses which, in turn, increases the dead-weight loss due to increased insurance costs and therefore is suboptimal.

THEOREM 2: $\bar{U}(P_0, I_0) \geq \bar{U}(P, I^*)$ *for all* $P \in S_2$.

The proof consists of comparing the slopes of the indifference curves for the insured and the insurer in $P, \bar{x}_2$ space. It is shown that for an incremental increase in $\bar{x}_2$ the insured is willing to increase P less than is required for the insurer to remain indifferent. Because of the limited space and since the proof is similar to the proof of Theorem 3 the details are omitted. The interested reader can receive the proof from the author upon request.

After showing that the Pareto optimal

insurance policy will not be of the upper-limit type, we now investigate the conditions under which the Pareto optimal policy will or will not include a deductible clause. The next theorem specifies that a (nontrivial) deductible is obtained if and only if the insurance cost depends on the insurance payment.

THEOREM 3: *A necessary and sufficient condition for the Pareto optimal deductible to be equal to zero is $c'(\cdot) \equiv 0$ (i.e., $c(I) = a$ for all I).*

The proof is given in Appendix B. We compare the insurer's and the insured's tradeoff between x_1 and P. If $c'(\cdot) = 0$, it is shown that for a marginal increase in the deductible the amount the insured is willing to pay in premium is less than that required by the insurer. On the other hand, if $c'(\cdot) > 0$, the insured is willing to pay more than what is required by the insurer and, therefore, the deductible is greater than zero.

Theorems 2 and 3 characterize the Pareto optimal insurance policy. What are the implications of these results regarding the contract form that we would expect to observe? The persistence of deductibles is explained by Theorem 3. If the cost of insurance depends on the coverage, then a nontrival deductible is obtained. This result does not depend on the risk preferences of the insured or the insurer. I stress this fact to point out that Arrow's (1971, Theorem 1) deductible result was not a consequence of the risk-neutrality assumption. Rather, it was obtained because of the assumption that insurance cost is proportional to coverage. For completeness Arrow's result is reproduced as a special case of my treatment.[3]

COROLLARY 1 (Arrow 1971): *If $c(I) = lI$ and the insurer is risk neutral, the Pareto optimal policy is given by*

$$(17) \qquad I^*(x) = \begin{cases} 0 & \text{for } x \leq \bar{x}_1 \\ x - \bar{x}_1 & \text{for } x > \bar{x}_1 \end{cases}$$

where $\bar{x}_1 > 0$ if and only if $l > 0$.

PROOF:

By Theorem 3, the Pareto optimal policy involves $\bar{x}_1 > 0$ iff $c' = l > 0$. The form of the coverage function is specified by (10). Since $R_V \equiv 0$ and $c'' = 0$, we have $I^{*\prime} = 1$ for $x > \bar{x}_1$. Thus $I^*(x)$ is given by (17).

Even if risk neutrality is not assumed, we still obtain a nontrivial deductible. The coverage involves, however, a coinsurance arrangement for losses above the deductible. The coinsurance level is given by (10) and in this special case is

$$I^{*\prime}(x) = \frac{R_U(A)}{R_U(A) + R_V(B)(1 + l)} < 1$$

This is the generalization of Arrow's (1971, Theorem 1) result to the risk-averse insurer case. Risk aversion of the insurer causes the coinsurance of losses above the deductible. Because of the insurance costs, the deductible is strictly positive.

What are the conditions that lead to a coinsurance arrangement? As already pointed out, risk aversion on the part of the insurer could be the cause for coinsurance. With no insurance costs this was proved by Arrow (1971, Theorem 2). In this case, there is no deductible. Our results prove, however, that a Pareto optimal policy may include a deductible and coinsurance of losses above the deductible.

COROLLARY 2: *If the insurer is risk averse, then the Pareto optimal insurance policy involves coinsurance of losses above the deductible.*

PROOF:

From (10) it is clear that $I^{*\prime} < 1$ for $x > \bar{x}_1 \geq 0$.

[3]Strictly speaking, Arrow (1971) did not fully discuss the Pareto optimal policy. His theorem characterizes only the optimal coverage function for a given premium.

Risk aversion, however, is not the only explanation for coinsurance. Even if the insurer is risk neutral, coinsurance might be observed, provided the insurance costs are a strictly convex function of the coverage. The intuitive reason for this result is that the cost function nonlinearity substitutes for the utility function nonlinearity.

COROLLARY 3: *If the insurer is risk neutral and* $c'' > 0$, *the Pareto optimal policy involves a deductible,* $\bar{x}_1 > 0$, *and coinsurance of losses above the deductible.*

PROOF:

Since $c' > 0$, we know $\bar{x}_1 > 0$ by Theorem 3. From (10) we have that for $x > \bar{x}_1$,

$$I^{*\prime}(x) = \frac{R_U(A)}{R_U(A) + c''/(1 + c')} < 1$$

In this section, the Pareto optimal insurance policy which was shown to specify a deductible and coinsurance of losses above the deductible was characterized. The deductible was shown to be strictly positive if and only if the insurance cost depended on the insurance payment. Coinsurance results from either insurer risk aversion or the cost function nonlinearity.

IV. Policies with Upper Limit on Coverage

The previous section explained deductibles and coinsurance arrangements in insurance contracts, as well as why Pareto optimal insurance policy does not involve an upper limit on coverage. In this section, I attempt to explain the prevalence of upper limits on coverage which are frequently incorporated in major medical, liability, and property insurance. The explanation rests on the fact that insurance companies are frequently regulated and therefore operate subject to a regulatory constraint. In what follows, it is argued that the upper limit on insurance coverage is desired by the insurance seller restricted in his policy offering by an actuarial constraint. Intuitively, the insurer is required to sell a policy with a prescribed actuarial value, for any given premium. This actuarial value might be smaller than the expected monetary loss, and then the policy cannot fully cover all potential losses. Being risk averse, the insurer prefers to allocate this given policy actuarial value to full coverage of small losses and limited coverage of large losses, rather than any other feasible form of the coverage. Heavy losses above that limit will not be insured under this contract.

To make the above statements precise, the results of the previous section are specialized to characterize the insurance policy desired by the insurance seller. Assume that the insurer devises contracts so as to maximize his expected utility of final wealth:

$$(18)\quad \bar{V}(P, I) = \max_{P, I(x)} \int_0^T V[W_0 + P - I(x) - c(I(x))] f(x)dx$$

The class of feasible insurance contracts is restricted by (1) and by the assumption that the premium received is required (by regulation) to be a function of the policy's actuarial value.[4] Denoting this function by R, assume

$$(19)\qquad P = R\left\{\int_0^T I(x) f(x)dx\right\}$$

Equation (19) specifies a general relationship between the premium charged and the actual value of the policy. This specification is consistent with the procedure used by regulatory agencies under the prior approval laws which are the predominant form of regulation of the property-liability insurance industry. (See Paul Joskow for the description of the pricing behavior in this industry.) In general, rates are established so as to yield a particular rate of return on sales (premiums). As Joskow states: "A standard rate of return on sales figure

[4]The result of this section would not change if we assume that the premium depends on the actuarial value of the coverage and insurance cost.

of 5 percent is employed in most states as a result of a recommendation by the National Association of Insurance Commissioners in 1921" (p. 394). Under this procedure, the pricing formula is

$$P(1 - .05) = \text{Expected Losses} + \text{Operating Expenses}$$

which is a special case of our formulation in (19).

Theorem 4 characterizes the solution to the insurer's problem specified above, stating that if a risk-averse insurer selects an insurance policy to maximize his expected utility, then the policy offered fully covers losses up to certain upper limit, and covers no losses above the limit.

THEOREM 4: *The solution to problem* (18) *subject to constraints* (1) *and* (19) *is* P^* *and* $I^*(x)$ *such that*

$$(20) \qquad I^*(x) = \begin{cases} x \text{ for } x \leq \bar{x} \\ \bar{x} \text{ for } x > \bar{x} \\ 0 \leq \bar{x} \leq T \end{cases}$$

and $\bar{x} = \bar{x}(P^*)$.

PROOF:

Starting with a fixed P, the optimal solution is shown to have the form of equation (20). The determination of the optimal P^* is then discussed.

The policies obtained in Theorem 1 can be viewed as the solution to the following problem: Maximize the insurer's expected utility of final wealth subject to constraint (1) and a restriction on the insured's expected utility of final wealth: $EU[w - P - x + I(x)] = c_1$, where c_1 is a constant. If the utility function U is linear and P is given, this restriction can be rewritten as $EI(x) = c_2$, which is equivalent to constraint (19). Thus, the solution to the present problem is obtained directly from Theorem 1 by specifying $R_U(\cdot) = 0$. This yields $I^{*\prime}(x) = 0$, for $x > \bar{x}_i$. In the first case, $I^*(x) = 0$ for $x \leq \bar{x}_1$, and, therefore, $I^*(x) = 0$ for all x. In the second case, $I^*(x) = x$ for $x \leq \bar{x}_2$, and, therefore, $I^*(x) = \bar{x}_2$ for $x > \bar{x}_2$ thus proving that the optimal form of the contract is given by (20). The constant $\bar{x}_2$ is determined by the optimal premium P^*, which depends on the function R and the insurance cost. If loading is sufficiently high, full coverage of all losses could be obtained (i.e., $\bar{x}_2 = T$). On the other hand, loading could be low enough so that no insurance is offered (i.e., $\bar{x}_2 = 0$). In general, therefore, $0 \leq \bar{x}_2 \leq T$.

V. Optimal Insurance Policies when Multiple Losses Can Occur

In the previous sections the insurance policy contracted between the insurance buyer and the insurance seller was analyzed. My model, however, incorporated the simplifying assumption that only a single loss can occur during the period of insurance protection. This assumption appears to be too restrictive; business firms and individuals may be faced with risks that could result in more than one loss during the period of insurance protection. Furthermore, the insurance buyer will typically purchase several different policies to cover different perils that he faces. The present analysis will extend the results of the previous sections to derive the properties of an optimal policy when several potential losses are faced by the insured. To facilitate notation, the proofs have been restricted to the case of two potential losses. The analysis carries over to more general cases.

The insurance buyer is assumed to face two potential losses during the period of insurance coverage. His total monetary loss is $x_1 + x_2$ where x_i, $i = 1, 2$, are assumed to be random variables defined on $[0, T_i]$ with a joint probability density function $f(x_1, x_2)$.

The insurance policy is characterized by the payment $I(x_1, x_2)$ transferred from the insurer to the insured if losses x_1, x_2 obtain. As before, $I(x_1, x_2)$ is referred to as the insurance policy or coverage function. Similar to condition (1) a restriction is imposed on the insurance function:

$$(21) \quad 0 \leq I(x_1, x_2) \leq x_1 + x_2 \qquad \text{for all } x_1, x_2$$

The Pareto optimal coverage function $I(x_1, x_2)$ is obtained by maximizing the insured's expected utility of final wealth subject to the constraint that the insurer's expected utility exceeds a given constant. The problem is then stated as follows:

$$(22) \quad \max_{P,I} \int_0^{T_2} \int_0^{T_1} U[w - P - x_1 - x_2 + I(x_1, x_2)] f(x_1, x_2) dx_1 dx_2$$

subject to

$$(23) \quad \int_0^{T_2} \int_0^{T_1} V[W_o + P - I(x_1, x_2) - c(I(x_1, x_2))] f(x_1, x_2) dx_1 dx_2 \geq k$$

and

$$(24) \qquad 0 \leq I(x_1, x_2) \leq x_1 + x_2$$

The above problem has the form of an iso-parametric problem in the calculus of variations with the additional constraint (24). Since the unknown function $I(x_1, x_2)$ depends on two variables, the extension of the simple Euler equation can be used to include two dimensions and constraints in order to derive the optimal insurance policy. Rather than proceeding along those lines, we first prove that the optimal function depends only on the sum $(x_1 + x_2)$.[5] Thus, the coverage function depends on one variable, the aggregate loss, and all previous results apply to this aggregate loss.

THEOREM 5: *Let $I^*(x_1, x_2)$ be the solution to the problem* (22) *subject to constraints* (23) *and* (24). *Then, $I^*(x_1, x_2)$ depends on the sum $(x_1 + x_2)$ only, i.e., $I^*(x_1, x_2) = \bar{I}^*(x_1 + x_2)$.*

The proof consists of showing that for any function $I(x_1, x_2)$ which does not depend on the sum only, there exists another coverage function $I^*(x_1, x_2)$ which increases the objective (22), is feasible, and depends on the sum only. The detailed proof is tedious and can be obtained from the author upon request. In what follows I provide the intuition behind the result and its proof.[6]

For simplicity, suppose that with probability $p(p')$ the losses are x_1, $x_2(x_1', x_2')$. Thus, the total loss is $x_1 + x_2$ or $x_1' + x_2'$ with probabilities p and p', respectively. Assume that $x_1 + x_2 = x_1' + x_2' = y$. Consider a coverage function $I(\cdot, \cdot)$ which does not depend only on the sum; $I(x_1, x_2) \neq I(x_1', x_2')$. A risk averter prefers to exchange any uncertainty for a certain outcome. In particular, the insured's expected utility for these two states can be increased by providing a coverage function which depends on y only:

$$pU[w - P - y + I(x_1, x_2)] + p'U[w - P - y + I(x_1', x_2')] \leq U[w - P - y + I^*(y)]$$

where $I^*(y) = pI(x_1, x_2) + p'I(x_1', x_2')$. Thus, the function I is dominated by the function I^*. I^* is the "weighted average" or the expected value of the coverages of equal total losses. By a similar argument, it can be shown that the insurer also prefers the coverage I^*. The above intuitive argument can be generalized. The driving force for the proof is, as above, the concavity of the utility functions; the dominance of I^* is established via Jensen's inequality.

Recall that our objective is to find optimal insurance policies when the insured is facing two potential losses. In Theorem 5 it was proved that any Pareto optimal coverage function depends only on the aggregate loss. The aggregate loss is denoted by $y = x_1 + x_2$ with probability density function $g(y)$. Using Theorem 5, we can rewrite problem (22)–(24) as: Find a coverage function $I(y)$ to maximize

$$\int_y U[w - P - y + I(y)] g(y) dy$$

subject to the constraints

[5] Borch (1962) proved that "any Pareto optimal set of treaties is equivalent to a pool arrangement" (p. 428). In his analysis, however, insurance was costless, and there were no constraints imposed on the feasible insurance policy. Therefore, he did not obtain the deductible or upper-limit results and could not generalize these results to the multiple loss case.

[6] I am indebted to Arie Tamir for suggesting this intuitive approach.

$$\int_y V[W_0 + P - I(y) - c(I(y))]g(y)dy \geq k$$
$$0 \leq I(y) \leq y$$

The above problem has the same structure as the problem considered in Sections II–IV with y replacing x. Thus, all the results regarding optimal insurance policies hold unchanged when the insured faces more than one risk, when the loss considered is the aggregate loss from all those risks. For example, if a risk-neutral insurer offers insurance policies and incurs linear cost, then it was proved that the Pareto optimal policy involves full coverage of losses beyond the deductible. Hence, we can now state:

COROLLARY 1': *If $c(I) = lI$ and the insurer is risk neutral, the Pareto optimal policy is given by:*

$$I^*(x_1, x_2) = \begin{cases} 0 & \text{for } x_1 + x_2 \leq \bar{x} \\ x_1 + x_2 - \bar{x} & \text{for } x_1 + x_2 > \bar{x} \end{cases}$$

where $\bar{x} > 0$ if and only if $l > 0$.

Similarly, all the theorems of the previous sections can be now restated with the only difference being that the loss considered is interpreted as the aggregate loss during the period of insurance protection.

VI. Conclusions

In this paper the prevalence of different insurance contracts was explained. It was shown that the Pareto optimal insurance contract involves a deductible and coinsurance of losses above the deductible. The deductible feature was shown to depend on the insurance costs. The coinsurance is due to either risk or cost sharing between the two parties. The upper limits on insurance were shown to be Pareto suboptimal. Their prevalence was shown to be in the interest of the regulated insurer. All results were obtained for single as well as multiple losses.

Two shortcomings of the above analysis should be noted. First, adverse selection problems were not analyzed; both the insurer and the insured were assumed to know the probability distribution function of the losses. Second, moral hazard problems were ignored; the monetary loss was assumed exogenous and not under the insured's control. A detailed analysis of the optimal contracts in these cases is much more difficult and was not attempted here.

Appendix A

PROOF of Lemma 2:

a) If $P \in S_1$, $I^*(x) = 0$ for $x \leq \bar{x}_1$. For $x > \bar{x}_1$, $I^*(x) = \int_{\bar{x}_1}^{x} I^{*\prime}(t)dt$, where $I^{*\prime}$ is given by (10). Differentiating with respect to $\bar{x}_1$,

$$\frac{\partial I^*(x)}{\partial \bar{x}_1} = -I^{*\prime}(\bar{x}_1) + \int_{\bar{x}_1}^{x} \frac{\partial I^{*\prime}(t)}{\partial I^*} \cdot \frac{\partial I^*(t)}{\partial \bar{x}_1} dt$$

Solving this equation yields

$$\frac{\partial I^*(x)}{\partial \bar{x}_1} = -I^*(\bar{x}_1) \exp\left\{\int_{\bar{x}_1}^{x} \frac{\partial I^{*\prime}(t)}{\partial I^*} dt\right\} < 0$$

Part (b) is proved similarly.

Appendix B

PROOF of Theorem 3:

From (7) we have that for all $P \in S_1$

$$\bar{V}(P, I^*) = \int_0^{\bar{x}_1} V[W_0 + P - a] f(x)dx + \int_{\bar{x}_1}^{T} V(B) f(x)dx = k$$

By differentiating we obtain expressions for $d\bar{P}/d\bar{x}_1$ when $\bar{V}$ and $\bar{U}$ are held constant. These are shown on page 95. From (12') and (14) we have that for $x \geq \bar{x}_1$

$$\text{(A1)} \quad \frac{U'(A)}{U'(w - P - \bar{x}_1)} = \frac{V'(B)(1 + c')}{V'(W_0 + P - a)[1 + c'(0)]}$$

Therefore,

$$\text{(A2)} \quad \frac{1}{U'(w - P - \bar{x}_1)} \int_{\bar{x}_1}^{T} U'(A) \frac{\partial I^*}{\partial \bar{x}_1} f(x)dx =$$

$$\frac{dP}{d\bar{x}_1}\bigg|_{\bar{V}=const} = \frac{\int_{\bar{x}_1}^{T} V'(B)(1+c')\frac{\partial I^*}{\partial \bar{x}_1} f(x)dx}{\int_0^{\bar{x}_1} V'[W_0+P-a]f(x)dx + \int_{\bar{x}_1}^{T} V'(B)[1-(1+c')\frac{\partial I^*}{\partial P}]f(x)dx}$$

$$\frac{dP}{d\bar{x}_1}\bigg|_{\bar{U}=const} = \frac{\int_{\bar{x}_1}^{T} U'(A)\frac{\partial I^*}{\partial \bar{x}_1} f(x)dx}{\int_0^{\bar{x}_1} U'(w-P-x)f(x)dx + \int_{\bar{x}_1}^{T} U'(A)\left(1-\frac{\partial I^*}{\partial P}\right)f(x)dx}$$

$$\frac{1}{V'(W_0+P-a)[1+c'(0)]} \int_{x_1}^{T} V'(B)(1+c')\frac{\partial I^*}{\partial \bar{x}_1} f(x)dx$$

To prove sufficiency, assume $c'(\cdot) = 0$. From (A1) and since x in the first integral is smaller than $\bar{x}_1$ we have

$$\text{(A3)}\quad \frac{1}{U'(w-P-\bar{x}_1)}\left\{\int_0^{\bar{x}_1} U'(w-P-x)f(x)dx + \int_{\bar{x}_1}^{T} U'(A)\left(1-\frac{\partial I^*}{\partial P}\right)f(x)dx\right\}$$
$$\leq \frac{1}{V'(W_0+P-a)}\left\{\int_0^{\bar{x}_1} V'(W_0+P-a)f(x)dx + \int_{\bar{x}_1}^{T} V'(B)\left(1-\frac{\partial I^*}{\partial P}\right)f(x)dx\right\}$$

Dividing (A2) by (A3) and recalling that, by Lemma 2, $\partial I^*/\partial \bar{x}_1 < 0$

$$\frac{dP}{d\bar{x}_1}\bigg|_{\bar{U}=const} \leq \frac{dP}{d\bar{x}_1}\bigg|_{\bar{V}=const}$$

with the equality holding only if $\bar{x}_1 = 0$. Thus the optimal policy is $\bar{x}_1 = 0$ and the premium is P_0 as was claimed.

To prove necessity, assume $c'(0) > 0$. There exists $y > 0$ such that

$$\frac{U'(w-P)}{U'(w-P-y)} = \frac{1}{1+c'(0)}$$

Using (A1) for $\bar{x}_1 > y$ we obtain

$$\text{(A4)}\quad \frac{1}{U'(w-P-\bar{x}_1)}\left\{\int_0^{\bar{x}_1} U'(w-P-x)f(x)dx + \int_{\bar{x}_1}^{T} U'(A)\left(1-\frac{\partial I^*}{\partial P}\right)f(x)dx\right\}$$
$$\geq \frac{1}{V'(W_0+P-a)[1+c'(0)]}\left\{\int_0^{\bar{x}_1} V'(W_0+P-a)f(x)dx + \int_{\bar{x}_1}^{T} V'(B)(1+c')\left(1-\frac{\partial I^*}{\partial P}\right)f(x)dx\right\}$$
$$> \frac{1}{V'(W_0+P-a)[1+c'(0)]}\left\{\int_0^{\bar{x}_1} V'(W_0+P-a)f(x)dx + \int_{\bar{x}_1}^{T} V'(B)\left[1-(1+c')\frac{\partial I^*}{\partial P}\right]f(x)dx\right\}$$

The last inequality is obtained since $c' > 0$. Dividing (A2) by (A4) and since $\partial I^*/\partial \bar{x}_1 < 0$ we have that for $\bar{x}_1 < y$

$$\frac{dP}{d\bar{x}_1}\bigg|_{\bar{U}=const} > \frac{dP}{d\bar{x}_1}\bigg|_{\bar{V}=const}$$

Therefore, the optimal deductible level in this case is different from zero, as we argued.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk Bearing*, Chicago 1971.

———, "Optimal Insurance and Generalized Deductibles," Rand Corp., R-1108-OEO, Feb. 1973.

K. Borch, "The Safety Loading of Reinsurance Premiums," *Skand. Aktuarietidskrift*, 1960, 162–84.

———, "Equilibrium in a Reinsurance Market," *Econometrica*, July 1962, *30*, 424–44.

J. P. Gould, "The Expected Utility Hypothesis and the Selection of Optimal Deductibles for a Given Insurance Policy," *J. Bus., Univ. Chicago*, Apr. 1969, *42*, 143–51.

P. L. Joskow, "Cartels, Competition and Regulation in the Property-Liability Insurance Industry," *Bell J. Econ.*, Autumn 1973, *4*, 375–427.

M. I. Kamien and N. L. Schwartz, "Sufficient Conditions in Optimal Control Theory," *J. Econ. Theory*, June 1971, *3*, 207–14.

J. Mossin, "Aspects of Rational Insurance Purchasing," *J. Polit. Econ.*, July/Aug. 1968, *76*, 533–68.

V. L. Smith, "Optimal Insurance Coverage," *J. Polit. Econ.*, Jan./Feb. 1968, *76*, 68–77.

R. B. Wilson, "The Theory of Syndicates," *Econometrica*, Jan. 1968, *36*, 119–32.

Income Redistribution: A Probabilistic Approach

By MICHAEL D. INTRILIGATOR*

The problem of the optimal distribution of income has been one of recurring interest to economists.[1] The approaches that have been taken generally involve either the implications of certain axiomatizations of the concepts of "equity" or "equality" or deductions from some general or specific social welfare function.

The purpose of this paper is threefold. First, it suggests that the problem be recast as one of income *re*distribution rather than one of income distribution, and it develops some principles for income redistribution (Section I). Second, it uses a probabilistic approach to social choice to generate a specific type of income distribution, using a linear income tax, that is consistent with the principles in the small-numbers case of a society with few individuals (Section II). Third, it generalizes the income distribution mechanism to a class of such mechanisms, the *linear income system*, analogous to the linear expenditure system of demand theory, and it shows that one member of this system satisfies a minimal incentive condition in the large-numbers case of a society with many individuals (Section III). Conclusions are summarized in Section IV.

I. Income Redistribution

Consider a society of m individuals, where individual i has income y_i; where he or she has the utility function

$$U_i = U_i(y_i) \tag{1}$$

which increases with income; and where each individual attempts to maximize utility.[2] In the customary approach to income distribution a (Bergson) *social welfare function*

$$W = W(U_1, U_2, \ldots, U_m) \tag{2}$$

is postulated, where social welfare is nondecreasing in each utility argument and where society attempts to maximize social welfare.[3] By adding more structure concerning the utility and welfare functions it is possible to reach certain conclusions regarding income distribution, such as the desirability of income equality.

A first purpose of this paper is to suggest that the problem be reformulated as one of income redistribution, that is, as the analysis of comparisons of the distribution of income before and after actual or hypothetical changes have occurred in this distribution. Such a reformulation would make the problem more meaningful and relevant, since in most societies the problem is, in fact, one of income redistribution rather than one of income distribution. In most societies, whether market oriented or socialist, there is no mechanism for creating an income distribution *de novo*. Rather there are usually mechanisms for influencing and modifying an existing income distribution, including taxes, social security, public assistance, rationing, price controls, provision of goods and services, expenditure policies, and many others. Thus the problem is one of starting from some pre-

*Professor of economics, University of California-Los Angeles. The research reported here was supported, in part, by a grant from the National Science Foundation. I would like to acknowledge with appreciation the valuable suggestions of D. L. Brito, James R. Meginniss, and an anonymous referee.

[1] Recent books on this subject include Amartya K. Sen (1973) and Anthony B. Atkinson (1975).

[2] This utility function can be interpreted as the indirect utility function, where the dependence on prices is implicit. On the indirect utility function, see the author (1971) and Louis Phlips.

[3] Two specific forms of the utility function, frequently discussed in the literature, for example, Sen (1977), are the *utilitarian* (*summation*) *form* $W_U = \Sigma U_i(y_i)$, where the sum ranges over all individuals, and the *Rawlesian* (*maximin*) *form* $W_R = \min U_i(y_i)$. See Sen (1970, 1973) for a critique of utilitarianism. For a more positive view see Peter Hammond. On the Rawlesian form see John Rawls and Edmund Phelps (1973b, 1977).

existing (actual or hypothetical) distribution of income and, by various combinations of these mechanisms, reaching a new income distribution, that is, a redistribution that is in some sense desirable.

The *initial distribution of income*, that is, the distribution of income before it is influenced by policy mechanisms, is given by the vector

$$\mathbf{y}^o = (y_1^o, y_2^o, \ldots, y_m^o) \tag{3}$$

where y_i^o is the initial income of individual i. The *final distribution of income*, that is, the distribution of income after it is influenced by policy mechanisms, is given by the vector

$$\mathbf{y} = (y_1, y_2, \ldots, y_m) \tag{4}$$

where y_i is the final income of individual i. Given the initial incomes, the problem of redistribution is that of choosing a set of effective taxes t_i (which represent the combined effect of all policy mechanisms). The initial income of individual i is reduced by the amount of the tax

$$y_i = y_i^o - t_i \tag{5}$$

but the tax can possibly be negative (in which case it represents a subsidy). The income redistribution is then summarized by the *tax vector*

$$\mathbf{t} = (t_1, t_2, \ldots, t_m) \tag{6}$$

The analysis of redistributions, as embodied in the tax vector (6), is in general not reducible to the analysis of distributions (4), such as that of the social welfare function (2), since the analysis of distributions generally does not identify the initial distribution of income in (3). Judging distribution by the social welfare function does not permit the initial distribution to be considered.[4] But the initial distribution is of critical importance in terms of the mechanisms typically available to a society to influence the distribution. In particular, taxes are levied on the initial distributions of income.

The tax vector is chosen subject to certain restrictions. First, aggregate income, obtained by summing income over all individuals, is in general not increased by the redistribution. Thus, if Y is total income after redistribution and Y^o is total income before redistribution[5]

$$Y \equiv \Sigma y_i \leq \Sigma y_i^o \equiv Y^o \tag{7}$$

so that from (5)

$$\Sigma t_i \geq 0 \tag{8}$$

The case where equality holds in (7) and (8) is one of *lump sum redistributions*, where the redistribution does not affect total income. Second, the final income for individual i should not fall below a *given base level of income* $\bar{y}_i$, which may be considered the subsistence level of income.[6] Thus

$$y_i \geq \bar{y}_i \tag{9}$$

so that from (5)

$$t_i \leq y_i^o - \bar{y}_i \tag{10}$$

It will be assumed that aggregate initial income exceeds the aggregate of base levels of income

$$Y^o \equiv \Sigma y_i^o > \Sigma \bar{y}_i \equiv \bar{Y}, \quad \text{or } Y^s \equiv Y^o - \bar{Y} > 0 \tag{11}$$

where total initial income Y^o less the total of the base levels of income, $\bar{Y}$ is Y^s, the *surplus income*, assumed positive. While aggregate initial income exceeds the aggregate base levels of income some individuals may have initial income less than their base income.

In general, income after redistribution for

[4]The social welfare function in (2) is "welfaristic" in the sense of Sen (1977) in that it uses no information about the social states other than the individual utility levels. Among other limitations, this welfaristic nature of the Bergson social welfare function rules out the use of information regarding the initial income distribution. A reformulation of (2) as an extended social welfare function

$$EW = EW(U_1, U_2, \ldots, U_m, y_1^o, y_2^o, \ldots, y_m^o) \tag{a}$$

could account for the initial income distribution.

[5]Here and elsewhere all sums range over all individuals, from 1 to m.

[6]The base level of income can be identified as the total expenditures on all base levels of goods and services in the linear expenditure system. For a discussion of this system see Phlips and the author (1977). Income tax laws generally identify such base levels of income by constructs such as the personal exemption, which at least originally had the intent of representing the minimum expenditures required for personal maintenance of the individual.

individual i can be considered a function of all initial levels of income and all base levels of income,[7]

$$(12)\quad y_i = f_i(y_1^o, y_2^o, \ldots, y_m^o; \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)$$

Thus, both initial and base levels of income are taken into account in determining taxes and the redistribution of income. With this formulation a reasonable redistribution should satisfy certain basic principles:

First Principle: Equally situated individuals, that is, individuals with the same initial level of income and with the same base level of income, should receive the same income. Thus, for all j, k,

$$(13)\quad y_j = y_k \quad \text{if} \quad y_j^o = y_k^o \quad \text{and} \quad \bar{y}_j = \bar{y}_k$$

This principle formalizes the concept of economic equality.

Second Principle: Each individual should receive a level of income no less than his/her base level, that is, for all i

$$(14)\quad y_i \geq \bar{y}_i$$

This principle ensures that no one in the society becomes destitute.

Third Principle: Each individual's income should always increase with his/her initial income. Thus for all i

$$(15)\quad \partial y_i / \partial y_i^o > 0$$

This principle ensures that all individuals have an incentive to earn more (initial) income.

Fourth Principle: Each individual's income should increase as total income increases, holding the individual's initial income and base income constant. Thus for all i

$$(16)\quad \partial y_i / \partial Y^o \mid_{y_i^o, \bar{y}_i \text{ constant}} > 0$$

This principle ensures that all individuals share in general increases in the level of well being of the society, even if their *own* initial income and base level of income do not change.

The next two sections will develop, respectively, one specific form of income redistribution and a general class of income redistributions that are consistent with some or all of these principles.

II. A Probabilistic Approach and the Equality Income System

The approach taken here to income redistribution will be based upon the probabilistic approach to social choice.[8] The connection to social choice is clear, since one of the most important issues of social choice is precisely that of income redistribution.

In the probabilistic approach to social choice each individual has a probability distribution over social outcomes, and the problem is that of aggregating these individual probabilities into a social probability distribution over the outcomes. The final choice of an outcome is then accomplished by a random mechanism using these probabilities. Under a reasonable set of axioms it has been shown that the social probabilities are simply the averages of the individual probabilities. More specifically, assume individual i would choose among the n alternatives for the society $A_1, A_2, \ldots, A_n$ according to the *individual probability vector*

$$(17)\quad \mathbf{q}_i = (q_{i1}, q_{i2}, \ldots, q_{in}), \quad i = 1, 2, \ldots, m$$

where q_{ij} is the probability that individual i would choose alternative A_j if he/she could act alone in deciding among the alternatives. These probability vectors could be, but are not restricted to, unit vectors. For example, if $\mathbf{q}_1 = (1/2, 1/2, 0, \ldots, 0)$ then individual 1 is indifferent between A_1 and A_2.

Let the *social probability vector* be

$$(18)\quad \mathbf{p} = (p_1, p_2, \ldots, p_n)$$

where p_j is the probability that the society chooses alternative A_j. The problem is then one of obtaining the social probability vector $\mathbf{p}$ from the individual probability vec-

[7]Each of the functions f_i is similar in form and interpretation to the allocation function introduced in Dagobert Brito, Anthony M. Buoncristiani, and the author, where the final allocation depends on both the initial allocations and the threat point.

[8]See the author (1973); see also Peter Fishburn.

tors $\mathbf{q}_i$. Three axioms are assumed: (i) *the existence of social probabilities* (i.e., given any set of m nonnegative vectors $\mathbf{q}_i$ with unit sums there exists a nonnegative vector $\mathbf{p}$ with a unit sum defined on the $\mathbf{q}_i$, so meaningful individual probabilities will yield meaningful social probabilities); (ii) *unanimity preserving for a loser* (i.e., if there is an alternative j such that $q_{ij} = 0$ for all i, then $p_j = 0$, so if all individuals choose a particular alternative with zero probability then so will society); and (iii) *strict and equal sensitivity of social probabilities to individual probabilities* (i.e., given an increase in any individual probability of choosing a particular alternative from q_{ij} to $q_{ij} + \Delta$, the social probability of that alternative will increase from p_j to $p_j + f(\Delta)$, where $f(\Delta) > 0$. Thus increasing any individual probability will increase the social probability for that alternative). Given these axioms there is a unique rule for determining social probabilities, the *average rule*, according to which the social probabilities are simple averages of individual probabilities[9]

$$(19) \qquad p_j = \frac{1}{m} \Sigma q_{ij}, j = 1, 2, \dots, n$$

In order to relate this result to the problem of income redistribution, it is necessary to specify the alternatives $A_1, A_2, \dots, A_m$ for the levels of income. Obviously there are many such alternative specifications. Guided by the four principles of income redistribution of Section I, one extreme alternative would be that in which there are as many alternatives as individuals ($m = n$) and in which alternative j is defined as

$$(20) \qquad A_j: y_j = \bar{y}_j + Y^s$$
$$\text{and} \qquad y_i = \bar{y}_i \quad \text{all } i \neq j$$

This is an extreme alternative for income redistribution since, while each individual receives his/her base level of income, the jth individual receives in addition the surplus income, that is, the total of all income in excess of the base levels of income. Thus everyone pays a tax equal to the excess of income over base income, and individual j receives all tax receipts. Equivalently, everyone other than j turns their income in excess of the base level of income over to individual j. Since utility in (1) is increasing in income, each individual would prefer the alternative in which he/she gains all of surplus income. Thus if individual i would act alone he/she would assign probability one to the alternative giving him/her the surplus income, i.e.,

$$(21) \qquad q_{ij} = \delta_{ij} = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} \text{ if } i \begin{Bmatrix} = \\ \neq \end{Bmatrix} j$$

where δ_{ij} is the Kronecker delta. Under the average rule (19) it follows that

$$(22) \qquad p_j = 1/m$$

implying that individual i has the probability $1/m$ of gaining the surplus income. This result is clearly one of equality in the sense that each individual has an equal chance of gaining all of surplus income.[10] The income of individual j after redistribution is

$$(23) \qquad y_j = \begin{Bmatrix} \bar{y}_j + Y^s \\ \bar{y}_j \end{Bmatrix} \text{with probability} \begin{Bmatrix} 1/m \\ 1 - 1/m \end{Bmatrix}$$

all individuals receiving their base level of income and individual j also receiving all of surplus income with probability $1/m$.

If all individuals are risk averse so that all the utility functions are strictly concave, then each individual would prefer to avoid the gamble represented by (23). Expected income for the gamble, however, is given as

$$(24) \qquad y_j^E = E(y_j) = y_j + \frac{1}{m} Y^s$$
$$= \bar{y}_j + \frac{1}{m}(Y - \bar{Y})$$

[9]For a proof and a comparison of the average rule to other rules, such as majority rule, the Borda rule, and the Pareto rule, see the author (1973).

[10]Put another way, after adjusting for base levels of income, each individual has an equal chance of being each person in the society with alternative income distributions representing alternative outcomes for the society. John Harsanyi and William Vickrey develop an expected utility approach to such situations.

and this level of income can in fact be paid to each individual. The resulting redistribution, called the ***equality income system***, is characterized by the equality of the income received by individuals over their base levels.[11] In the lump sum redistribution case, where total income is the same before and after redistribution, the tax yielding the equality income system can be written as the following function of initial (pretax) income y_j^o

$$(25)\qquad t_j^E = \left(1 - \frac{1}{m}\right) y_j^o - \left[\frac{1}{m}\sum_{i \neq j} (y_i^o - \bar{y}_i) + \left(1 - \frac{1}{m}\right)\bar{y}_j\right]$$

Thus it is a linear function of pretax income, as illustrated in Figure 1.

Several important characteristics should be noted for the linear income tax given in (25) and illustrated in Figure 1. First, the marginal tax rate is constant at all levels of income and is given by

$$(26)\qquad \partial t_j^E / \partial y_j^o = 1 - 1/m$$

The marginal tax rate is therefore identical for all individuals. Second, the negative intercept shown in Figure 1 implies that individuals with low pretax income will receive a subsidy, shown as the dotted portion of the line and corresponding precisely to a negative income tax.[12] The maximum subsidy at zero pretax income is given by the term in square brackets in (25), representing the proportionate share of the surplus of all *other* individuals and the base level of the individual in question, the latter weighted by the residual weight $1 - 1/m$. This income, which is the lowest level of income received after redistribution, exceeds the base level of income. It is shown in Figure 1 as the negative of the intercept on the t_j^E axis. Third, the individual receives a subsidy if pretax income lies below the cutoff level shown as $\hat{y}_j^E$ in Figure 1, while he/she pays a tax if pretax income lies above the cutoff level. This cutoff level is given as

$$(27)\qquad \hat{y}_j^E = \bar{y}_j + \frac{1}{m-1}\sum_{i \neq j} (y_i^o - \bar{y}_i)$$

that is, as the base level plus the average over all $m - 1$ *other* individuals of their incomes in excess of the base levels. Since this cutoff level of pretax income exceeds the base level, individual j receives a subsidy automatically if his/her initial income lies below the base level, and even individuals with pretax income in excess of the base level may receive a subsidy. Fourth, the tax is a progressive one in that the

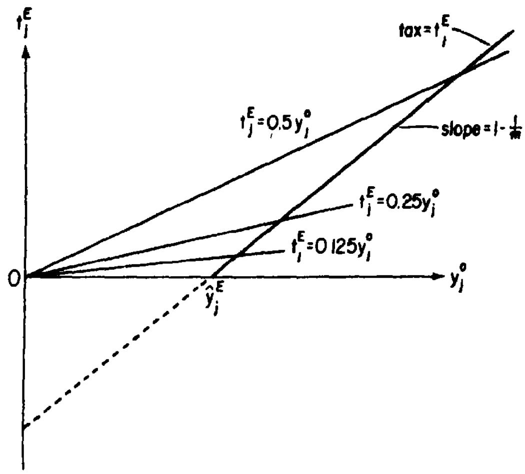


FIGURE 1. TAX PAID AS A FUNCTION OF PRETAX INCOME FOR THE EQUALITY INCOME SYSTEM (SEE (25) AND (27), RESPECTIVELY, FOR DEFINITIONS OF t_j^E AND $\hat{y}_j^E$)

[11] Hammond proves that this system is optimal if there is "complete dual interpersonal comparability," that is, equity judgments are based on levels of the utility function and the social welfare function is of the utilitarian form given in fn. 3.

[12] Both James Mirrlees and Robert Aumann and Mordecai Kurz proved that the optimal income tax involves such a negative income tax at low levels of income. Eytan Sheshinski (1972) proved that the optimal linear income tax provides a positive lump sum at zero income, as here, assuming the supply of labor is a nondecreasing function of the net wage and leisure is a normal good. He also proved that the optimal linear income tax uses a marginal tax rate that is bounded above by $1/(1 + \lambda)$, where λ is the lowest elasticity of the labor supply function. This result is consistent with the marginal tax rate in (26) if this lowest elasticity is less than $1/(m - 1)$, which, for practical purposes, is met if the lowest elasticity is zero. See also Phelps (1973a), Efraim Sadka, Sheshinski (1977) and Brito and William Oakland.

proportion of income paid as a tax rises with income, as shown by the rays from the origin, corresponding to 1/8, 1/4, and 1/2 of income, respectively.

The equality income system, summarized by the income after redistribution (24) or, equivalently, by the tax function (25), provides a specific example of a redistribution system satisfying the four principles of Section I in the small-numbers case of a society with few individuals. First, equally situated individuals are treated equally. Second, all individuals receive more than their base level of income. Third, income always increases with initial income, where, assuming lump sum redistribution,

$$\partial y_j^E/\partial y_j^o = 1/m > 0 \tag{28}$$

Fourth, income rises as total income increases, where

$$\partial y_j^E/\partial Y^o\Big|_{y_j^o,\bar{y}_j\,constant} = 1/m > 0 \tag{29}$$

Now consider the large-numbers case when m becomes very large, such as a nation with tens or hundreds of millions of individuals. Then the derivatives in (28) and (29) become vanishingly small. In this case there is no perceptible effect of either an increase in initial income or an increase in total income on the income of individual i. In fact, in the limit as $m \to \infty$ income is given as

(30)

$$\lim y_j^E = \bar{y}_j + \frac{1}{m-1}\sum_{i\neq j}(y_i^o - \bar{y}_i) = \hat{y}_j^E$$

so that each individual receives exactly the cutoff level of income $\hat{y}_j^E$. In effect, all income is confiscated and then redistributed, with everyone receiving the base level plus the average of the total surplus of all other individuals' incomes. Thus the posttax level of income is independent of the pretax income, violating the third principle of income redistribution. Such a system raises a problem of incentives in this large-numbers case, specifically the lack of incentives for an individual to earn (pretax) income, given that his/her posttax income is independent of whatever they personally earn. When m is small, however, the equality income system is consistent with all the principles and is quite workable. It may be a reasonable system of income redistribution in a family, where each individual receives a base level of income (for example, a weekly allowance) plus an equal share of the surplus over total base levels (for example, the family takes a vacation together). It may also be a reasonable system for a business partnership such as a law firm, where each partner receives a base level of income (for example, a specified salary) plus an equal share of the surplus over total base levels (for example, an end-of-year bonus).

The next section introduces a generalization of the equality income system representing a class of redistributions. This class is of interest in its own right, but in addition, it provides possible redistributions that can overcome the incentive problem encountered in the equality income system in the large-numbers case.

III. The Linear Income System and the Proportional Linear Income System

A general class of income redistribution systems, of which the equality income system is one member, is the *linear income system*. In this system income after redistribution is

$$y_j^* = \bar{y}_j + \beta_j Y^s = \bar{y}_j + \beta_j(Y - \bar{Y}) \tag{31}$$

where each individual receives his/her base income plus a share of the surplus income. The system is defined by the set of base levels of income $\bar{y}_j$ and the shares β_j, and it is similar in specification and interpretation to the linear expenditure system of demand theory, justifying the choice of the name "linear income system."[13] The shares

[13] In the linear expenditure system, as discussed in the references in fn. 6, y_j^* in (31) would be interpreted as the expenditure by one consumer on good j, (rather than the income of individual j), $\bar{y}_j$ would be interpreted as the base expenditure on good j (rather than the base income of individual j), and Y^s would be interpreted as "supernumerary income," that is, income of individual j in excess of the total of all base expenditures (rather than total income in excess of all base incomes). Thus the linear income system can be considered a way of redistributing income

β_j are called ***marginal income shares***, by analogy to the "marginal budget shares" of the linear expenditure system, and they satisfy[14]

(32) $\beta_j \geq 0,\ \Sigma\beta_j = 1$

The equality income system of Section II is such a linear income system in which the marginal income shares are all equal

(33) $\beta_j = 1/m$

More generally, the marginal income shares reflect the relative "deservingness" of the individuals in the society. In fact, if β_j is interpreted as the probability that individual j receives all surplus income then y_j^* in (31) is simply the expected income of individual j.

The marginal income shares can, in general, be functions themselves of the initial and base levels of income[15]

(34) $\beta_j = \beta_j(y_1^o, y_2^o, \ldots, y_m^o; \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)$

The tax function for the linear income system is

(35) $t_j^* = (1 - \beta_j)y_j^o - [\beta_j \sum_{i \neq j}(y_i^o - \bar{y}_i) + (1 - \beta_j)\bar{y}_j]$

among individuals in a society that is similar in form to the way the linear expenditure system allocates income for one individual among all goods.

[14]Hammond develops a system similar in form to the linear income system in the case of "incomplete dual comparability," in which equity judgments are based on levels of the utility function and the social welfare function is of the form $W = \Sigma\beta_i U_i(y_i)$. Hammond suggests, as one interpretation of the β_i, assuming i refers to a household rather than an individual, the number of "adult equivalents" in household i relative to the total number of adult equivalents in all households. A specific social welfare function of this form which generates the linear income system is that using the logarithmic utility functions $W = \Sigma\beta_i \ln(y_i - \bar{y}_i)$. Maximizing this W by choice of lump sum redistributions yields (31), just as maximizing a utility function of this form yields the linear expenditure system. See also Atkinson (1970) and David Champernowne.

[15]Note that the absence of any system of redistribution can be interpreted as a linear income system where the marginal income shares satisfy $\beta_j = (y_j^o - \bar{y}_j)/Y^s$ representing the proportion of surplus income contributed by individual j. In such a status quo system $y_j = y_j^o$. Of course this system does not satisfy the second principle, since individuals need not receive their base levels of income.

Thus the marginal tax rate is given as

(36) $\dfrac{\partial t_j^*}{\partial y_j^o} = (1 - \beta_j) - \dfrac{\partial \beta_j}{\partial y_j^o} Y^s$

where the second term on the right allows for the influence of the initial level of income on the marginal income shares. For this system, then, the marginal tax rate will generally vary with income.

An example of the linear income system in which the marginal income shares vary with income is the *proportional income system*. In this system the marginal income shares are the proportion of total income contributed by the individual

(37) $\beta_j = y_j^o / Y$

so that "deservingness" increases with income. Assuming lump sum redistribution the tax function is

(38) $t_j^p = \left(\dfrac{\bar{Y}}{Y}\right) y_j^o - \bar{y}_j$

where $\bar{Y}/Y$ is the ratio of total base income to total income. The marginal tax rate is

(39) $\dfrac{\partial t_j^p}{\partial y_j^o} = \left(1 - \dfrac{y_j^o}{Y}\right)\dfrac{\bar{Y}}{Y}$

Income is subject to tax (subsidy) if it exceeds (is less than) the cutoff level

(40) $\hat{y}_j^p = \bar{y}_j\left(\dfrac{Y}{\bar{Y}}\right)$

that is, base income scaled up by the ratio of total income to total base income, a ratio which can be considered a measure of the well being of the society. Since

(41) $Y/\bar{Y} = 1/(1 - Y^s/\bar{Y})$

in a society with little surplus income this ratio is close to unity, so the cutoff level of income is close to the base level. By contrast, in a society with large surplus income the ratio exceeds unity, so the cutoff level exceeds the base level. For example if surplus income is half of income then the ratio is 2, so income is subject to taxation only if it exceeds twice the base level of income.

In the proportional income system income is given by

(42)

$$y_j^p = \bar{y}_j + y_j^o \frac{Y^s}{Y} = y_j^o + \left(\bar{y}_j - y_j^o \frac{\bar{Y}}{Y}\right)$$

Income is therefore base income plus initial income scaled down by the ratio of surplus income to total income, or equivalently, initial income plus the excess of base income over initial income scaled by the ratio of total base income to total income.

The proportional income system is one that satisfies all the principles of income redistribution.[16] First, from (42) individuals with the same initial income and base income receive the same income. Second, everyone receives a level of income no less than his/her base income. In fact, from (42) individuals receive at least the base income $\bar{y}_j$, and they receive more as y_j^o increases. Third, everyone's income increases with initial income, where, in the large-numbers case

$$\lim \partial y_j^p / \partial y_j^o = Y^s / Y > 0 \tag{43}$$

In terms of taxes, in the large-numbers case

$$\lim \frac{\partial t_j^p}{\partial y_j^o} = \frac{\bar{Y}}{Y} < 1 \tag{44}$$

so even in this limiting case marginal tax rates are less than one (and all are equal). Fourth, if income Y increases, where base levels of income remain constant (so the increase in Y^s is equal to the increase in Y) then

$$\left.\frac{\partial y_j^p}{\partial Y^o}\right|_{y_j^o, \bar{y}_j \text{ constant}} = \frac{y_j^o}{Y} \frac{\bar{Y}}{Y} > 0 \tag{45}$$

The main difference between these results and those for the equality income system involves the third principle. In the equality income system added initial income does not increase income in the large-numbers case, while in the proportional income system case added initial income always does increase income, even in the large-numbers case.

[16]Of course the proportional linear income system is not the only one to satisfy all the principles. In fact any linear income system for which $\beta_j = f_j(y_j^o)/\Sigma f_j(y_j^o)$, where $f_j(y_j^o)$ is any nonnegative function of initial income, will also satisfy all the principles. This system can be considered a *generalized proportional income system*.

IV. Conclusions

This paper has emphasized income redistribution as opposed to the more customary emphasis on income distribution, and it has introduced four principles of income redistribution. By taking into effect both initial levels of income (before redistribution) and base levels of income (for example, subsistence levels) a general class of redistributions, the linear income system, has been formulated. This system is analogous to the linear expenditure system of demand theory, income being given by the base levels of income plus the marginal income share times the surplus income (i.e., income in excess of total base income). In general the marginal income shares can depend on all initial and base levels of income.

One special case, the equality income system, uses constant and equal marginal income shares. This system satisfies all the principles of income redistribution in the small-numbers case of a society with a small number of individuals, for example, a family or a business partnership, and it may be reasonable in such a setting. It fails, however, to satisfy the principle that income increases with initial income in the large-numbers case.

Another special case, the proportional income system, uses marginal income shares equal to the proportion of initial income contributed by the individual. This system satisfies all the principles of income redistribution in the small-numbers case, and, unlike the equality income system, it satisfies the principle that income increases with initial income even in the large-numbers case. This system, or one like it, may be a reasonable one when there are large numbers of individuals among whom income is to be redistributed.

REFERENCES

Anthony B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

———, *The Economics of Inequality*, Oxford 1975.

R. J. **Aumann and M. Kurz**, "Power and Taxes," *Econometrica*, July 1977, *45*, 1137–61.

D. L. Brito, A. M. Buoncristiani, and M. D. Intriligator, "A New Approach to the Nash Bargaining Problem," *Econometrica*, July 1977, *45*, 1163–72.

——— **and W. H. Oakland**, "Some Properties of the Optimal Income Tax," *Int. Econ. Rev.*, June 1977, *18*, 407–23.

David G. Champernowne, *The Distribution of Income between Persons*, Cambridge 1973.

P. C. Fishburn, "A Probabilistic Model of Social Choice: Comment," *Rev. Econ. Stud.*, Apr. 1975, *42*, 297–301.

P. J. Hammond, "Dual Interpersonal Comparisons of Utility and the Welfare Economics of Income Distribution," *J. Publ. Econ.*, Feb. 1977, *7*, 51–71.

J. C. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *J. Polit. Econ.*, Aug. 1955, *63*, 309–21.

Michael D. Intriligator, *Mathematical Optimization and Economic Theory*, Englewood Cliffs 1971.

———, "A Probabilistic Model of Social Choice," *Rev. Econ. Stud.*, Oct. 1973, *40*, 553–60.

———, *Econometric Models, Techniques, and Applications*. Englewood Cliffs; Amsterdam 1978.

J. A. Mirrlees, "An Exploration in the Theory of Optimal Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, 175–208.

E. S. Phelps, (1973a) "Taxation of Wage Income for Economic Justice," *Quart J. Econ.*, Aug. 1973, *87*, 331–54.

———, (1973b) *Economic Justice*, Hammondsworth 1973.

———, "Recent Developments in Welfare Economics: Justice et Equité," in Michael D. Intriligator, ed., *Frontiers of Quantitative Economics*, Vol. III, Amsterdam 1977.

Louis Phlips, *Applied Consumption Analysis*, Amsterdam 1974.

John Rawls, *A Theory of Justice*, Cambridge 1971

E. Sadka, "On Income Distribution, Incentive Effects, and Optimal Income Taxation," *Rev. Econ. Stud.*, June 1976, *43*, 261–7.

Amartya K. Sen, *Collective Choice and Social Welfare*, San Francisco 1970.

———, *On Economic Inequality*, Oxford 1973.

———, "On Weights and Measures: Informational Constraints in Social Welfare Analysis," *Econometrica*, Oct. 1977, *45*, 1539–72.

E. Sheshinski, "The Optimal Linear Income Tax," *Rev. Econ. Stud.*, July 1972, *39*, 297–302.

———, "Income Inequality and Growth," in Michael D. Intriligator, ed., *Frontiers of Quantitative Economics*, Vol. III, Amsterdam 1977.

W. S. Vickrey, "Utility, Strategy, and Social Decision Rules," *Quart. J. Econ.*, Nov. 1960, *74*, 507–35.

A Theoretical Foundation for the Gravity Equation

By James E. Anderson*

Probably the most successful empirical trade device of the last twenty-five years is the gravity equation. Applied to a wide variety of goods and factors moving over regional and national borders under differing circumstances, it usually produces a good fit. Unfortunately, as is widely recognized, its use for policy is severely hampered by its "unidentified" properties. Insertion into the equation of policy instruments such as border taxes has no theoretical justification; and inference about the effect of taxes from examining changes in the equation over times when taxes have changed carries no guarantee of validity.

The gravity equation ordinarily is specified as

$$(1) \qquad M_{ijk} = \alpha_k Y_i^{\beta_k} Y_j^{\gamma_k} N_i^{\xi_k} N_j^{\epsilon_k} d_{ij}^{\mu_k} U_{ijk}$$

where M_{ijk} is the dollar flow of good or factor k from country or region i to country or region j, Y_i and Y_j are incomes in i and j, N_i and N_j are population in i and j, and d_{ij} is the distance between countries (regions) i and j. The U_{ijk} is a lognormally distributed error term with $E(\ln U_{ijk}) = 0$. Frequently the flows are aggregated across goods. Ordinarily the equation is run on cross-section data and sometimes on pooled data. Typical estimates find income elasticities not significantly different from one and significantly different from zero, and population elasticities around $-.4$ usually significantly different from zero.[1]

The intent of this paper is to provide a theoretical explanation for the gravity equation applied to commodities. It uses the properties of expenditure systems with a maintained hypothesis of identical homothetic preferences across regions. Products are differentiated by place of origin (for a justification, see Peter Isard). The gravity model constrains the pure expenditure system by specifying that the share of national expenditure accounted for by spending on tradeables (openness to trade) is a stable unidentified reduced-form function of income and population. The share of total tradeable goods expenditure accounted for by each tradeable good category across regions is an identified (through preferences) function of transit cost variables. Partial identification is achieved. While other interpretations are possible (see for example Edward Leamer and Robert Stern),[2] the one advanced here has

*Professor of economics, Boston College. I am indebted to Marvin Kraus and Edward Leamer for helpful comments.

[1]See for example Norman D. Aitken.

[2]They offer three explanations. The first, based on physics, has little interest. The second identifies the equation loosely as a reduced form with exogenous demand-side variables (importer income and population) and supply-side variables (exporter income and population). Alternatively, the importer and exporter characteristics identify the size of the foreign sector in each, with any flow a function of size at either end. The third interpretation is based on a probability model. Let Z_i be country i's total imports, an unidentified reduced-form function of income, population, and other possibly unobservable variables. The set $\{Z_i/T\}$, where $T = \Sigma_i Z_i$, world trade, has the form of a probability distribution. Alternatively Z_i/T is a trade potential. The probability of the occurrence of flow between i and j is taken to be $Z_i Z_j/T^2$. Alternatively, potential between i and j is the product of the i and j potentials. The expected size of the flow given T is then $M_{ij} = Z_i Z_j/T$. The term T is constant in a cross-section study and can be neglected. Resistance to trade, proxied by distance, can be inserted and with the *log*-linear form for all functions, we have the gravity equation. This interpretation has the advantage of explaining the multiplicative functional form, and has a useful flexibility. Leamer subsequently developed a hybrid version of it to explain aggregate imports of good k by country i. In the hybrid model, Z_j becomes $Z_{j(i)} = \Sigma_{j \neq i} Z_j$. Also, the parameters of the Z_i and $Z_{j(i)}$ functions are permitted to vary by commodity group k. Leamer's hybrid is thus

$$M_{ik} = Z_i^k Z_{j(i)}^k \Psi(\mathbf{d}_i, \mathbf{t}_i, \mathbf{t}_{j(i)})$$

where $\mathbf{d}_i$ is a vector of distance from i to all other countries, $\mathbf{t}_i$ is a vector of i's tariffs, and $\mathbf{t}_{j(i)}$ is a vector of all other countries' tariffs. The problem with either the Leamer-Stern gravity interpretation or the Leamer hybrid is that while the potential or probabil-

four distinct advantages. First, it explains the multiplicative form of the equation. Second, it permits an interpretation of distance in the equation, identifying the estimated coefficient, and can be used as part of an attack on estimating the effect of instrument changes. Third, the vague underlying assumption of identical "structure" across regions or countries is straightforwardly interpreted as identical expenditure functions. This suggests appropriate disaggregation. Finally, following the logic of the present interpretation implies that the usual estimator of the gravity equation may be biased, requiring change in the method of estimation.

The present interpretation of the gravity model makes it part of an alternative method of doing cross-section budget studies. The bias problems now uncovered may be quite severe, especially with transit costs varying considerably, but there are efficiency gains to trade off against them. The background of difficulty in modelling trade flows requires respect for any potentially promising method. This paper shows that the gravity model may merit continued development and use.

Section I develops the simplest linear expenditure model, which produces an equation like (1) but with the last three variables omitted and with Y_i and Y_j constrained to have unit elasticity. The major portion of the explanatory power of the gravity model is thus encompassed. While yielding a gravity equation, the models would never sensibly be so estimated.

In the next two sections, the gravity approach gains legitimacy as a device offering large gains in efficiency of estimation at a possible cost of bias. An important fact of life is large interregional and international variations in shares of total expenditure accounted for by traded goods, even across regions or countries where spending patterns are reasonably similar (for example, the set of developed Western countries). These are assumed to vary as a function of national income and population. Total trade expenditure is distributed across individual categories by share functions which are identical across countries. With this structure, Section II produces a gravity equation with potentially attractive properties. Section III discusses estimation of the model, shows that the usual technique may produce biased results, and suggests alternatives. Section IV integrates in distance (as a proxy for transport costs) and border taxes producing a full model suggesting the possibility of identifying long-run tariff elasticities. A constant elasticity of substitution (*CES*) case developed in the Appendix provides further details.

Two areas for theoretical development may be noted. The major remaining unidentified part of the equation is the function stipulating that trade's share of budgets is dependent on income and population. While this is a well-established empirical relation, it would be nice to have an explanation. None is offered here.

Use of pooled cross-section and time-series data requires a further development of the model also not attempted here. Two requirements should be noted. First, a theory of how short-run responses to price changes (revealed over time) are related to long-run responses (revealed over the cross section) is needed. Second, a theory of how short-run responses to income changes (revealed over time in the Keynesian type of trade model) are related to long-run responses must be constructed. Part of the second story must be the relation of trade balance to asset accumulation stressed in the recent monetary approach to the balance of payments.

I. The Pure Expenditure System Model

The simplest possible gravity-type model stems from a rearrangement of a Cobb-Douglas expenditure system. Assume that each country is completely specialized in the production of its own good (as in a Keynesian-type trade model), so there is one good for each country. No tariffs or transport costs exist. The fraction of income spent on the product of country i is

ity story is plausible, it lacks a compelling economic justification.

denoted b_i and is the same in all countries (i.e., there are identical Cobb-Douglas preferences everywhere). With cross-section analysis, prices are constant at equilibrium values and units are chosen such that they are all unity. Consumption in value and quantity terms of good i in country j (= imports of good i by country j) is thus

$$M_{ij} = b_i Y_j \tag{2}$$

where Y_j is income in country j.

The requirement that income must equal sales implies that

$$Y_i = b_i(\Sigma_j Y_j) \tag{3}$$

Solving (3) for b_i and substituting into (2), we obtain

$$M_{ij} = Y_i Y_j / \Sigma Y_j \tag{4}$$

This is the simplest form of "gravity" model. If we disregard error structure,[3] a generalization of equation (4) can be estimated by ordinary least squares, with exponents on Y_i, Y_j unrestricted. In a pure cross section, the denominator is an irrelevant scale term. The income elasticities produced (disregarding bias) should not differ significantly from unity. The functional form of the gravity equation and a major portion of the explanatory power is encompassed by the expenditure system model.

II. The Trade-Share-Expenditure System Model

The gravity equation of Section I is based on identical Cobb-Douglas preferences, implying identical expenditure shares and gravity equation income elasticities of unity. It could be fancied up by allowing policy induced price differences to produce different expenditure shares in a less restrictive preference form such as the *CES*, but there is little point in the exercise. While a gravity equation is produced by such a framework, the real variables of interest are the non-income-dependent expenditure shares. The gravity equation is a silly specification from an econometric standpoint since it substitutes out the share (which in the Cobb-Douglas case is the only parameter). This section appends to the Cobb-Douglas expenditure system for traded goods a differing traded-nontraded goods split and produces an unrestricted (non-unit income elasticity) gravity equation. The next section shows that the gravity equation becomes far more sensible.

Traded-goods shares of total expenditure vary widely across regions and countries. Hollis Chenery and others subsequently have found that in cross-section data such shares are "explained" rather well by income and population. Moreover, the linear or *log*-linear regression line of traded goods shares on income and population tends to be stable over time. No identification of this relationship is attempted here, but loosely, income per capita is an exogenous demand-side factor, and population (country size) a supply-side factor. Trade shares "should" increase with income per capita and decrease with size. Leamer and Stern have also suggested including an endowment measure as an explanatory variable which would act somewhat like size.[4] Accepting the stability of the trade-share function, the expenditure system model combines with it to produce the gravity equation.

Assume that all countries produce a traded and a nontraded good. The overall preference function assumed in this formu-

[3]Note that the manipulation which justified the presence of Y_i in (4) means that there may be simultaneous equation bias, with the Y's not being independent of error terms postulated for the equations (2) and (3). This problem is discussed in Section III.

[4]It is easy to construct examples where the trade share is a simple closed form function of factor endowment variables. Consider the small-country case in which all traded goods may be treated as a composite and there is one nontraded good. Stipulate a simple non-Cobb-Douglas utility function in traded goods and the nontraded good and assume the linear version of the two-sector production model. When a model of this sort is solved for an interior equilibrium-traded-goods share assuming trade balance, it produces a relatively simple function of the endowments. Results analogous to the Johnson pro- and antitrade bias analysis are embodied in the function. I have been unable, however, to discover a specification which produces a traded-goods share function which is *log*-linear in income and factor endowments.

tion is weakly separable with respect to a partition between traded and nontraded goods: $u = u$ (g(traded goods), nontraded goods). Then given the level of expenditure on traded goods, individual traded-goods demands are determined as if a homothetic utility function in traded goods alone $g(\ )$ were maximized subject to a budget constraint involving the level of expenditure on traded goods. The individual traded-goods shares of total trade expenditure with homotheticity are functions of traded-goods prices only.[5] For simplicity, it is assumed $g(\ )$ has the Cobb-Douglas form in the rest of the text. Within the class of traded goods, since preferences are identical, expenditure shares for any good are identical across countries. Thus, for any consuming country j, θ_i is the expenditure on country i's tradeable good divided by total expenditure in j on tradeables; i.e., θ_i is an exponent of $g(\ )$. Let ϕ_j be the share of expenditure on all traded goods in total expenditure of country j and $\phi_j = F(Y_j, N_j)$.

Demand for i's tradeable good in country j (j's imports of i's good) is

$$(5) \qquad M_{ij} = \theta_i \phi_j Y_j$$

The balance-of-trade relation for country i implies

$$(6) \qquad Y_i \phi_i = \Big(\sum_j Y_j \phi_j\Big)\theta_i$$

value of imports of i plus domestic spending on domestic tradeables	=	value of exports of i plus domestic spending on domestic tradeables

Solving (6) for θ_i and substituting into (5), we have

$$(7) \qquad M_{ij} = \frac{\phi_i Y_i \phi_j Y_j}{\sum_j \phi_j Y_j} = \frac{\phi_i Y_i \phi_j Y_j}{\sum_i \sum_j M_{ij}}$$

With $F(Y_i, N_i)$ taking on a *log*-linear form, (7) is the deterministic form of the gravity equation (1) with the distance term suppressed and a scale term appended. More realistically, if trade imbalance due to long-term capital account transactions is a function of (Y_i, N_i), we may write the "basic" balance $Y_i \phi_i m_i = (\Sigma_j Y_j \phi_j)\theta_i$, with $m_i = m(Y_i, N_i)$, and substitute into (6) and (7). This yields[6]

$$(8) \qquad M_{ij} = \frac{m_i \phi_i Y_i \phi_j Y_j}{\sum_i \sum_j M_{ij}}$$

With *log*-linear forms for m and F, (8) is again essentially the deterministic gravity equation.

III. Estimation Efficiency

The model of linear expenditures of Section I, while implying a gravity equation, would never sensibly be so estimated. Homothetic preferences identical across countries imply identical expenditure share functions, and these can be estimated directly, treating distance and trade taxes appropriately in a manner set out in Section IV. Simply divide the demand equations of Section I or their analogue by own income and find the mean over j of M_{ij}/Y_j = the estimator of θ_i. The stochastic budget constraint information can also be utilized to (in effect) add one observation since[7] $\theta_i = Y_i/\Sigma_j Y_j$.

The trade-share model of Section II on the other hand lends some legitimacy to the gravity model. Eventually we will allow

[5]Homotheticity of $g(\ )$ is imposed because the presence of traded-goods expenditure as an argument in the θ_i function will greatly complicate estimation. Separability is imposed to permit the two-stage decision process which removes nontraded-goods prices from the θ_i function. Some justification for separability may be found in the observation that traded goods are far more similar to each other than they are to nontraded goods. Homotheticity would imply that, *ceteris paribus*, larger nations' trade expenditures are scalar expansions of smaller nations' trade expenditures. This may not do great violence to reality. A gravity-type model which imposes neither restriction is possible, but it would be far more complex and difficult to estimate. As with most empirical work on preferences, the restrictions' main appeal is convenience.

[6]Balance-of-payments disequilibrium could be treated as part of the error terms in estimation. Use of the model is best restricted to equilibrium years, however, since error terms due to disequilibrium may be correlated with Y_i, N_i, and therefore cause errors in estimation.

[7]Again, we disregard the problem that the Y's depend on the error term through the budget constraint.

many tradeables for each country, with tariffs and transport costs present, but initially, as before, assume only one tradeable in each and no barriers to trade. The system to be estimated is

$$(5') \qquad M_{ij} = \Theta_i \phi_j Y_j U_{ij}$$

$$(6') \qquad m_i \phi_i Y_i = \Theta_i \sum_j \phi_j Y_j$$

where U_{ij} is a *log*-normal disturbance with $E(\ln U_{ij}) = 0$. Note that (6′) states that planned expenditures (reduced or increased by the capital account factor) = planned sales, and has no error term. Efficient estimation requires that the information in (6′) be utilized. The most convenient way to do this, since the constraint is highly nonlinear in the Y's, is to substitute out Θ_i and estimate the gravity equation:

$$(8) \quad M_{ij} = \frac{m(Y_i, N_i)F(Y_i, N_i)Y_i F(Y_j, N_j)Y_j}{\sum_j F(Y_j, N_j)Y_j} U_{ij}$$

With the *log*-linear form for $m(\ )$ and $F(\ )$,

$$m(Y_i, N_i) = k_m Y_i^{m_y} N_i^{m_N}$$

and $$F(Y_i, N_n) = k_\phi H_j^{\phi_y} N_j^{\phi_N}$$

and the denominator made a constant term we have

$$(8') \quad M_{ij} = (k_m Y_i^{m_y} N_i^{m_N})(k_\phi Y_i^{\phi_y} N_i^{\phi_N}) Y_i \cdot (k_\phi Y_j^{\phi_y} N_j^{\phi_N}) Y_j U_{ij} + k'$$
$$= (k_m k_\phi^2) Y_i^{m_y + \phi_y + 1} N_i^{m_N + \phi_N} \cdot Y_j^{\phi_y + 1} N_j^{\phi_N} U_{ij} + k'$$

This is the aggregate form of (1) with the distance term omitted. Ordinarily it would be fitted on a subset of countries in the world. Exports to the rest of the world are exogenous and imports from it are excluded from the fitting. When this is done, the denominator is still the sum of world trade expenditures, and (6′) implies that (8) and (8′) assume that Θ_i is the same in the excluded countries as in the included countries. Alternatively, (6′) can be interpreted as a payments union multilateral balance constraint (which includes as a special case the rest of the world account being always zero). The denominator of (8) and (8′) then has only the included group's trade expenditure.[8] Under either interpretation, the identifying restrictions immediately allow recovery of all structural exponents from the estimator of (8′). Either interpretation will also permit the complex constant term $(k_m k_\phi^2/k')$ to be unravelled, though the estimators $\hat{k}_m, \hat{k}_\phi, \hat{k}'$ have only large sample unbiasedness.[9] Finally, form the set of estimated values for traded-goods expenditures:

$$(9) \qquad \hat{\phi}_j Y_j = \hat{k}_\phi Y_j^{\hat{\phi}_y + 1} N_j^{\hat{\phi}_N}$$

The individual traded-goods shares Θ_i can be estimated using the instruments $\hat{\phi}_j Y_j$ (which are asymptotically uncorrelated with U_{ij}):

[8]If neither of these alternatives is palatable, exports to the rest of the world can be fixed at $M_{i,n+1}$. Trade balance is now

$$(a) \qquad m_i \phi_i Y_i = \Theta_i \sum_j \phi_j Y_j + M_{i,n+1}$$

When the trade balance is solved for Θ_i and substituted, the gravity equation is

$$(b) \qquad M_{ij} = \frac{(m_i \phi_i Y_i - M_{i,n+1})\phi_j Y_j}{\sum_j \phi_j Y_j}$$

Non-linear methods must be used to estimate (b).

[9]Consider the worldwide identity of preferences case Using conditional expectations in (6′), the trade balance requirement implies that

$$k_m k_\phi Y^{1+\phi_y+m_y} N_i^{\phi_N+m_N} = \sum_{j=1}^{n} E(M_{ij}) + M_{i,n+1}$$

where $M_{i,n+1}$ is the exogenous nonrandom export of i to the rest of the world. Replacing $E(M_{ij})$ with its solved values $\hat{M}_{ij}$, and the exponents with their estimated values, we have

(a)

$$\hat{k}_m \hat{k}_\phi = [\sum_{j=1}^{n} \hat{M}_{ij} + M_{i,n+1}] \, Y_i^{-(1+\hat{\phi}_y+\hat{m}_y)} N_i^{-(\hat{\phi}_N+\hat{m}_N)}$$

Using the definition of k', we have

$$(b) \qquad \hat{k}' = \hat{k}_\phi(\sum_{j=1}^{n} Y_j^{1+\hat{\phi}_y} N_j^{\hat{\phi}_N}) + \sum_{i=1}^{n} M_{i,n+1}$$

Finally the estimated constant term k, is theoretically related to the three constants $\hat{k}_m, \hat{k}_\phi, \hat{k}'$ by

$$(c) \qquad \hat{k} = \hat{k}_m \hat{k}^2/\hat{k}'$$

(a)–(c) can be solved explicitly for the three constants $\hat{k}_m, \hat{k}_\phi, \hat{k}'$.

$$M_{ij} = \hat{\theta}_i \phi_j Y_j U_{ij} \tag{10}$$

which is estimated across countries for country i's exports (including the rest of the world's exports to included countries), subject to the restriction that $\Sigma\theta_i = 1$. The alternative without the gravity model is to estimate the ϕ's by regressing $\Sigma_i M_{ij}/Y_j$ on $F(Y_j, N_j)$ and then repeating the second stage. The gravity equation in effect squares the number of observations used in estimating the parameters of $F(Y, N)$. For the limited number of cross-section observations available, the gain in efficiency should be large.

It is ironic, however, that the very simultaneity which allows substitution for θ_i may imply that the Y's are mutually determined with the error terms of the expenditure system. The model (5′)-(6′) postulates no random term in the trade balance constraint and thus allows treatment of the Y's as predetermined. Suppose alternatively that the trade balance constraint is a stochastic form:

$$m_i(\sum_j M_{ji}) = \sum_j M_{ij} \tag{6''}$$

or $$m_i \phi_i Y_i(\sum_i \theta_j U_{ji}) = \theta_i(\sum_j \phi_j Y_j U_{ij})$$

Then (8) becomes

$$M_{ij} = \frac{m_i \phi_i Y_i \phi_j Y_j}{\sum_j \phi_j Y_j} \epsilon_{ij} \tag{8''}$$

where $$\epsilon_{ij} = U_{ij} \sum_j \theta_j U_{ji} / \sum_j \frac{\phi_j Y_j}{\sum_j \phi_j Y_j} U_{ij}$$

A regression based on (8″) with $m(\ )$ and $F(\ )$ assigned the multiplicative form will produce biased results (with unknown direction) due to the dependency of the Y's on the error terms. The relative stability of the equation over time in some applications may suggest that the bias is not serious, but this is conjectural. Several alternatives are possible, two of which will be discussed here. Something like the gravity equation would be desirable, replacing the Y's in the equation with the instruments highly correlated with the Y's but independent of the demand equation error terms. One such instrument might be lagged income, particularly for years when last year's income seems unlikely to be correlated with this year's error term. Bias remains, but may be reduced. The other alternative is to attempt dealing with simultaneity directly. Suppose a subset of countries is considered and preferences for traded goods are everywhere the same. Rest of the world demand is considered exogenous. Run the gravity equation using ordinary least squares on the *log* of the equation (or non-linear estimation of the equation with an additive disturbance term) and obtain estimates of ϕ_i, $\hat{\phi}_i$, and m_i, $\hat{m}_i$; $i = 1, \ldots, n$.[10] The trade balance equations in matrix notation are

$$(diag\ \hat{m})(diag\ \hat{\phi})\, Y = (\hat{\theta}\iota')(diag\ \hat{\phi}) \cdot Y + M_{n+1} \tag{11}$$

where M_{n+1} = rest of the world demand, an $n \times 1$ vector
Y = $n \times 1$ vector of incomes
$(diag\ \hat{\phi})$ = $n \times n$ diagonal matrix with $\hat{\phi}_i$, $i = 1, \ldots, n$ on the diagonal
$\hat{\theta}$ = $n \times 1$ vector of $\hat{\theta}_i$, $i = 1, \ldots, n$
ι' = $1 \times n$ row vector of ones
$(diag\ \hat{m})$ = $n \times n$ diagonal matrix with $\hat{m}_i$, $i = 1, \ldots, n$ on the diagonal

or[11] $$Y = (diag\ \hat{\phi})^{-1}[I - \hat{\theta}\iota']^{-1} M_{n+1}$$

The left-hand side contains instruments for the Y's which attempt to deal with the simultaneity problem, and which can then be inserted into the gravity equation and used to reestimate the m's, ϕ's, and θ's. Continued iteration would have no necessarily desirable property.

Either instrument would probably be preferable if the simultaneity problem were severe, as in the European Economic Community (*EEC*). For groups of countries with

[10]Since $E(ln\ \epsilon_{ij}) \neq 0$, previous methods of identifying k_m and k_ϕ cannot be used. For simplicity, this detail can be evaded by assuming $k_m = k_\phi = 1$.

[11]Provided $\Sigma_i \theta_i < 1$, $I - \theta\iota'$ has an inverse. We deal with a subset of goods and countries, so the condition is fulfilled. $M_{n+1} = \theta\phi_{n+1} Y_{n+1}$ with the assumption of identical preferences for the rest of the world.

relatively small interdependence, the gravity equation with the Y's directly used might be preferable, the greater efficiency of direct use of Y's dominating the bias. These are, of course, only rules of thumb.[12]

IV. Many Goods, Tariffs, and Distance

Now consider the gravity equation under the complication of many commodity classes of goods flowing between each country i and j, with a full set of national tariffs in each country, and with transport costs proxied by distance. Preferences for traded goods are identical across countries and are homothetic, with the traded-goods share, as before, a function of income and population. Within each commodity class, goods are considered to be differentiated by place of origin.[13] The gravity equation still has use in the estimation of trade-flow equations of this system. As in Section III, it is a device for increasing the efficiency of estimation of the trade-share function parameters. Unfortunately, tariffs and transport costs create added sources of bias in estimation of both stages. The gain may still outweigh the loss.

The landed value at country j of commodity class k goods produced in country i is $M_{ijk}\tau_{ijk}$, where M_{ijk} is the foreign port value and τ_{ijk} is the transit cost factor (including all border adjustments and transport costs). With identical homothetic preferences for traded goods, the traded-goods expenditure shares are identical functions $\Theta_{ik}(\tau_j)$, where τ_j is the vector of the τ_{ijk}'s for country j. Demand for import ik (with foreign port prices of unity as before) is

$$(12) \qquad M_{ijk} = \frac{1}{\tau_{ijk}} \Theta_{ik}(\tau_j)\phi_j Y_j$$

Aggregate trade flows between i and j are thus

$$(13) \qquad M_{ij} = \sum_k M_{ijk} = \phi_j Y_j \sum_k \frac{1}{\tau_{ijk}} \Theta_{ik}(\tau_j)$$

The trade balance relation is

$$(14) \qquad m_i\phi_i Y_i = \sum_j M_{ij} = \sum_j \phi_j Y_j \sum_k \frac{1}{\tau_{ijk}} \Theta_{ik}(\tau_j)$$

Previously we set all the $\tau_{ijk} = 1$ and could divide both sides of (14) by $\Sigma_j \phi_j Y_j$ to obtain the aggregate share parameter for country i goods on the right: $\Sigma_k \Theta_{ik}$. The left-hand side was then substituted into (13) to obtain the gravity equation

$$(8) \qquad M_{ij} = \frac{m_i\phi_i Y_i \phi_j Y_j}{\sum_j \phi_j Y_j}$$

Note that with many goods, only the aggregate version of the gravity equation is valid under the present interpretation.[14]

[12] We should note that in principle minimum-distance or full-information maximum-likelihood techniques can be used to estimate the parameters of system (5′)-(6′) and its generalization. Any model which can be solved to isolate its disturbance terms can be so treated (see A. R. Gallant). The costliness and convergence difficulties with such non-linear techniques makes compromises like those in the text attractive.

[13] Isard offers an empirical justification for this assumption. On a theoretical level, note that the gravity model almost necessarily implies differentiation by place of origin. How else can (i) two-way flows be explained, and (ii) the flow of good k between points i and j be modelled as a function of variables at i and j alone?

[14] Leamer has extended the gravity-type cross-section model to aggregation across partner countries j, estimating aggregate outward flows M_{ik}. His model shares with the present interpretation unidentified reduced-form trade potential functions similar to the trade-share functions above. It is further unidentified because Leamer gives no precise economic reason for the appearance and form of the appearance in the equation of trade potential at both ends of the trade flow. Nevertheless, it has a certain plausibility and allows extension to the estimation of tariff elasticities. In correspondence concerning an earlier version of this paper, Leamer suggested a method of obtaining commodity-specific gravity equations not involving the trade potential interpretation. As before, the demand function is

$$(a) \qquad M_{ijk} = \Theta_{ik}\phi_j Y_j$$

Country i's income from sales of the ik good is

$$(b) \qquad Y_{ik} = \Theta_{ik}\sum_j \phi_j Y_j$$

Suppose that the commodity classes k are defined so that income from their production across countries

With the τ_{ijk} departing from unity the division of both sides of (14) by $\Sigma_j \phi_j Y_j$ produces

$$(15)\quad \frac{m_i \phi_i Y_i}{\sum_j \phi_j Y_j} = \sum_j \frac{\phi_j Y_j}{\sum_j \phi_j Y_j} \cdot \sum_k \frac{1}{\tau_{ijk}} \Theta_{ik}(\tau_j)$$

The gravity equation substitutes for the share in (13), $\Sigma_k(1/\tau_{ijk})\Theta_{ik}(\tau_j)$, a weighted average of such shares across all countries j. This will cause bias of unknown sign in the gravity equation parameter estimator based on the stochastic version of (12)–(14), and subsequently in the parameter estimator of the demand equation (12).[15] Other factors being equal, the bias will be less the more closely the transit costs resemble one another. In the limit, we return to the model of Section III. Evidently, similarity of transit costs should be a criterion for selecting countries in the cross-section sample. This is too stringent a criterion to permit the viability of the gravity model and flies in the face of the role it assigns to distance. To breathe life back into it, we can argue that with dissimilarity of restricted types it may still be possible to escape with small bias.

If transit costs of all sorts are an increasing function of distance and the same across commodities ($\tau_{ijk} = f(d_{ij})$ with $f(0) = 1$ and $f' > 0$), then with Cobb-Douglas preferences the demand equation and trade balance equations are

$$(13')\quad M_{ij} = (\sum_k \Theta_{ik})\phi_j Y_j \frac{1}{f(d_{ij})} U_{ij}$$

$$(15')\quad m_i \phi_i Y_i = (\sum_k \Theta_{ik}) \sum_j \phi_j Y_j \frac{1}{f(d_{ij})}$$

Equation (13') states that the foreign port value of country j's demand for all of i's goods equals country j's total expenditure on traded goods (in home prices), $\phi_j Y_j$, times the common aggregate traded-goods expenditure share for i's goods $\Sigma_k \Theta_{ik}$ deflated by the transit cost factor. Equation (15') states that country i's expenditure on all traded goods at i's prices $\phi_i Y_i$ times the capital account scale factor m_i must equal the value at country i of i's exports to all countries. The gravity equation can now be derived as

$$(16)\quad M_{ij} = \frac{m_i \phi_i Y_i \phi_j Y_j}{\sum_j \phi_j Y_j} \cdot \frac{1}{f(d_{ij})} \cdot \left[\sum_j \frac{\phi_j Y_j}{\sum_j \phi_j Y_j} \cdot \frac{1}{f(d_{ij})}\right]^{-1} U_{ij}$$

With m and the ϕ's made *log*-linear functions of income and population, (16) resembles (1), with three differences. First, (16) is an aggregate equation rather than commodity specific. Second, $1/f(d_{ij})$ is not a *log*-linear function.[16] Finally, the square bracket term is missing in (1). It can be interpreted as saying that the flow from i to j depends on economic distance from i to j relative to a trade-weighted average of economic distance from i to all points in the system. The model leading to (16) is probably the best case one can make for the aggregate gravity equation as it is usually fitted in practice. The square bracket term might have little variation across origin points i for a group of countries distributed geographically in a polygon (for example, the *EEC*). Changing origin point i will lengthen some distances and shorten others, with the potential for little change in the weighted average. With small enough bias,

is a stable function of *GNP*, population and resource endowments E_i:

$$(c)\quad Y_{ik} = \gamma^k(Y_i, N_i, E_i) Y_i$$

Substituting (b) and (c) into (a) we obtain a gravity form

$$(d)\quad M_{ijk} = \frac{\gamma^k(Y_i, N_i, E_i) Y_i \phi_j(Y_j, N_j, E_j) Y_j}{\sum_j \phi_j Y_j}$$

This may be a promising approach, although the stability of the γ^k functions is probably more controversial than the stability of the ϕ_j functions.

[15]The demand equation as before would be estimated using $\hat{\phi}_j Y_j$ as an instrument. Note that the parameters to be estimated would in principle include substitution parameters.

[16]Practitioners of the gravity model use it only because it is so convenient, and some have adopted more theoretically appealing forms.

the greater efficiency of the gravity equation (1) in aggregate form in arriving at estimates of $\phi_j Y_j$ dominates,[17] and the Θ's can be estimated from

$$(12') \quad M_{ijk} = [(\widehat{\frac{1}{f(d_{ij})}})(\hat{\phi}_j Y_j)]\, \Theta_{ik} U_{ijk}$$

If the bias from omitting the bracketed term is likely to be substantial, (16) can be estimated with constant weights in the bracket term equal to observed trade total expenditure shares. Non-linear least squares is required, implying some loss in efficiency. Which procedure is preferable depends on the tradeoff.

Practitioners may be able to get away with restrictions less extreme than either the identical transit costs or Cobb-Douglas assumptions. Consider the *CES* preference case where trade taxes are the same across all countries j for any good k of country i (often an assumption which comes close to reality, as in the *EEC*). Assume transport cost factors depend only on distance (not on commodity group). Under these conditions, the Appendix shows that a gravity equation may still have some promise of providing efficiency gains which dominate bias. The *CES* demand functions may be estimated as above in a second stage using the instrument $\phi_j \hat{Y}_j$. In principle, a trade flow system in the gravity model style capable of dealing with tariffs and even possibly with policy-induced change in shares can be developed.

V. Conclusion

The gravity equation can be derived from the properties of expenditure systems. In this interpretation it is an alternative method of doing cross-section budget studies, and one with potentially important efficiency properties. Its use is at the widest limited to countries where the structure of traded-goods preference is very similar and, subsidiarily, where trade tax structures and transport cost structures are similar. In future work, it would be desirable to learn more about the tradeoff between bias and efficiency involved in the gravity equation. Other extensions include building an intertemporal version and identifying the trade-share function.

Appendix: The *CES* Case

The *CES* traded goods utility indicator is

$$U_j = [\sum_i \sum_k \beta_{ik} M_{ijk}^{-\rho}]^{-1/\rho}$$

where M_{ijk} is the quantity of good k from country i consumed in country j. Good k is a different commodity in each country due to differentiation. The elasticity of substitution is $\sigma = 1/(1+\rho)$. Expenditure shares derived from such a utility function are

$$(A1) \qquad \Theta_{ijk} = \frac{\beta_{ik}(P_{ijk})^{1-\sigma}}{\sum_i \sum_k \beta_{ik}\sigma(P_{ijk})^{1-\sigma}}$$

where Θ_{ijk} is the traded-goods expenditure share of country j for good k of country i, and P_{ijk} is the price of good k from country i landed in country j. The denominator of (A1), when raised to the power $1/(1-\sigma)$, gives the "true cost-of-living" index for the *CES* function. The demand for imports is

$$(A2) \qquad M_{ijk} = \Theta_{ijk}\phi_j Y_j \frac{1}{P_{ijk}}$$

Derivations are standard, so omitted.

Assume now that the transit cost factors

[17] Disregarding bias, the procedure for solving out the parameters of the $m(\ )$ and $F(\ )$ functions are essentially the same as in Section II. The only new factor is the presence of $f(d_{ij})$. If we adopt the *log*-linear form, (16) assures us that any constant term it possesses is cancelled out (i.e., if $f(d_{ij}) = k_d d_{ij}^{\mu}$, the k_d term would not appear in the general constant term of the estimator of (16)). The constant term k_d can be identified by noting that

$$(e) \qquad \frac{\hat{M}_{ij}}{\hat{M}_{ii}} = \frac{Y_j^{\hat{\phi}_y+1} N_j^{\hat{\phi}_N}}{Y_i^{\hat{\phi}_y+1} N_i^{\hat{\phi}_N}} \frac{1}{k_d d_{ij}^{\hat{\mu}}}$$

Equation (a) of fn. 9 becomes

$$(a') \quad \hat{k}_m \hat{k}_\phi \hat{k}_d = [\sum_{j=1}^{\hat{}} M_{ij} + \hat{M}_{i,n+1}]\, Y_i^{-(1+\hat{\phi}_y+\hat{m}_y)} N_i^{-(\hat{\phi}_N+\hat{m}_N)} d_{ij}^{-\hat{\mu}}$$

Equations (a'), (e), as well as (b) and (c) of fn. 9 can be solved for all constant term estimates. Other distance functions will require other identifying restrictions.

are based on two components. The first is t_{ik}, a tax on good k of country i levied by all countries in the group. The second is a transit cost factor, common to all goods and dependent on distance, $h(d_{ij})$.

$$\text{(A3)} \qquad \frac{P_{ijk}}{P_{ik}} = t_{ik} h(d_{ij})$$

Define the "free trade" share:

$$\text{(A4)} \qquad \Theta_{ik} \equiv \frac{\beta_{ik} P_{ik}^{1-\sigma}}{\sum_i \sum_k \beta_{ik}^{\sigma} P_{ik}^{1-\sigma}}$$

Using (A3) and (A4) in (A1) and (A2) the demand equation can be written

$$\text{(A5)} \quad M_{ijk} = \Theta_{ik} \left[\frac{\dfrac{[t_{ik} h(d_{ij})]^{1-\sigma}}{\sum_i \sum_k \beta_{ik}^{\sigma} P_{ik}^{1-\sigma}(t_{ik} h(d_{ij}))^{1-\sigma}}}{\sum_i \sum_k \beta_{ik}^{\sigma} P_{ik}^{1-\sigma}} \right] \cdot \phi_j Y_j \frac{1}{P_{ik} t_{ik} h(d_{ij})}$$

The denominator of the large square bracket term is a weighted average of the transit cost factors, and equals a transit cost true cost-of-living index for country j raised to the power $1 - \sigma$. Denote this as $g_j^{1-\sigma}$. Simplifying (A5) and using the convention that free trade prices are unity:

$$\text{(A5}') \quad M_{ijk} = \Theta_{ik} g_j^{-(1-\sigma)} (t_{ik} h(d_{ij}))^{-\sigma} \phi_j Y_j$$

The trade balance requirements are

$$\text{(A6)} \quad m_i \phi_i Y_i = \sum_j \sum_k \Theta_{ik} g_j^{-(1-\sigma)} \cdot (t_{ik} h(d_{ij}))^{-\sigma} \phi_j Y_j$$

Aggregate trade flows between i and j are

$$\text{(A7)} \quad M_{ij} = \phi_j Y_j g_j^{-(1-\sigma)} \; h(d_{ij})^{-\sigma} \sum_k \Theta_{ik} t_{ik}^{-\sigma}$$

The gravity equation substitution replaces $\Sigma_k \Theta_{ik} t_{ik}^{-\sigma}$ in (A7) with $m_i \phi_i Y_i / \Sigma_j \phi_j Y_j$. The proper substitution yields

$$\text{(A8)} \quad M_{ij} = \left[\frac{g_j^{-(1-\sigma)}}{\sum_j \dfrac{\phi_j Y_j}{\sum_j \phi_j Y_j} g_j^{-(1-\sigma)} h(d_{ij})^{-\sigma}} \right] \cdot \frac{m_i \phi_i Y_i \phi_j Y_j}{\sum_j \phi_j Y_j} [h(d_{ij})]^{-\sigma}$$

The gravity equation run on the stochastic version of (A8) omitting the square bracket term has a chance of reasonably small bias in the estimator if there is little variation in the square bracket term as we move across i and j. Note that with free trade prices of unity:

$$\text{(A9)} \quad g_j^{1-\sigma} = \frac{\sum_i \sum_k \beta_{ik}^{\sigma} (t_{ik})^{1-\sigma} h(d_{ij})^{1-\sigma}}{\sum_i \sum_k \beta_{ik}^{\sigma}}$$

The denominator of the square bracket is a weighted sum of $h(d_{ij})^{-\sigma}$ across j for a given i. The numerator is a weighted sum of $h(d_{ij})^{-\sigma}$ across i for a given j. Changing origin points i and destination points j may well create changes which wash out (as in the related square bracket term of (16) in the text). This is more likely for countries geographically distributed in a polygon (and impossible for countries distributed on a line). We might thus hope to gain more in efficiency than we lose in bias by estimating the stochastic form of (A8) omitting the bracket term. The parameters of the $m(\ )$ and $F(\ )$ functions can be identified as in the text, and the instruments $\hat{\phi}_j Y_j$ used to attack the stochastic form of the demand equation (A5'). Note that in estimating (A5') we might again appeal to the lack of variation in $g_j^{1-\sigma}$ to produce simpler estimation techniques capable of producing asymptotically "unbiased" estimates of σ.

Other alternatives include approximation of g_j in the numerator of (A8) with a Laspeyres traded-goods price index, and full non-linear estimation of the stochastic form of (A8).

REFERENCES

N. D. Aitken, "The Effect of the *EEC* and *EFTA* on European Trade: A Temporal Cross-Section Analysis," *Amer. Econ. Rev.*, Dec. 1973, *63*, 881–92.

H. B. Chenery, "Patterns of Industrial Growth," *Amer. Econ. Rev.*, Sept. 1960, *50*, 624–54.

A. R. Gallant, "Three Stage Least Squares Estimation for a System of Simultaneous,

Non-Linear, Implicit Equations," *J. Econometrics*, Feb. 1977, *5*, 71–88.

P. Isard, "How Far Can We Push The Law of One Price?," *Amer. Econ. Rev.*, Dec. 1977, *67*, 942–48.

Edward E. Leamer and Robert M. Stern, *Quantitative International Economics*, Boston 1970.

———, "The Commodity Composition of International Trade in Manufactures: An Empirical Analysis," *Oxford Econ. Pap.*, Nov. 1974, *26*, 350–74.

A Model of the Natural Rate of Unemployment

By STEVEN C. SALOP*

Since the publication of Edmund Phelps' volume, the "new" macroeconomics has treated the labor market as a dynamic process of rational search by unemployed workers for available vacancies. Wages are viewed as at least potentially flexible, though free contracting between workers and firms may lead to fixed wages in the short run. Imperfect information is a crucial element of the theory, for it implies both a need for contracting and a need for rational search rather than simple market clearing in each period.

A positive rate of frictional unemployment may exist in equilibrium, denoted as the "natural" rate. This unemployment is due to the frictions in the search process and imperfections in information rather than to any deficiency in aggregate demand. Milton Friedman defined the natural rate as

> the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is embedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on. [p.8]

This paper reexamines the micro foundations of the natural rate in a model of labor market equilibrium in which turnover flows and imperfect information are explicitly considered. Workers may quit their current jobs to enter the unemployment pool in order to search among available vacancies for a more preferred position. Firms economize on turnover by an appropriate wage policy. The model to be presented is essentially a stationary analogue to models formulated by Dale Mortensen and Phelps (1970b), with one major difference. In this model, the internal labor market for experienced trained workers is conceptually separated from the external labor market for new employees. Moreover, the firm is constrained by morale, moral hazard, and capital market imperfections to pay an identical wage rate to all its employees, regardless of seniority. As a result, both labor markets are unable to clear simultaneously, and in general, quantity adjustments are required in one of the markets. I focus on the case in which the quantity adjustments take place in the external new applicant market.

As a result of this friction in the labor market, equilibrium entails not only the usual voluntary frictional component of unemployment, but possibly also a component of involuntary unemployment. This involuntary unemployment is permanent; it may not be eliminated through aggregate monetary or fiscal policy. Instead it is structural in the sense that it derives from the inability of all markets to clear simultaneously, a friction that is imbedded in the structure of the economy. The equilibrium also contains components of disguised unemployment and search unemployment.

I. The Model[1]

The formal model has the following basic structure. The labor market contains no uncertainty in the aggregate, though every

*Federal Trade Commission. The remarks in this statement represent only my personal views They are not intended to be, and should not be construed as, representative of the views of any other member of the Federal Trade Commission staff or individual Commissioners. This paper is dedicated to Al Klevorick, who convinced me to fully complete my dissertation with this paper and Edmund Phelps, who originally stimulated these ideas. David Soskice rekindled my interest in the problem and Dale Mortensen has provided continuing encouragement I am grateful to George Akerlof, Steve Salant, Joseph Stiglitz, and the referee for helpful comments and insights, and Mary Ann Henry for superb typing and editing.

[1]This section follows the author (1973b).

worker and firm does face some private uncertainty. When a new employee joins a firm, he is uncertain of the particular set of nonpecuniary characteristics offered by the firm, but learns them through experience on the job. Once these characteristics become known, if the employee is dissatisfied and believes he can do better elsewhere, he quits and joins the unemployment pool to search for alternative employment. (In order to keep the model simple, on-the-job search is ignored.) Quits depend on the tightness of the labor market, rising when unemployment is low and falling when opportunities are scarcer. Unemployment and wage rates adjust until the costs of turnover to firms and the benefits of quitting to workers are equilibrated.

Turnover is costly to firms through its direct costs such as formal orientation programs, expenditures to foremen for "breaking in" new employees as well as indirect costs such as lowered productivity during the adjustment process. As a result, firms utilize wage policy to economize on turnover. This concern for turnover occurs regardless of conditions in the external labor market. Even if a lost worker can be immediately replaced with an identical new applicant, the new applicant is less valuable than an experienced worker, since the turnover costs must be borne again.

Since experienced workers are more valuable to the firm, we would expect to observe wage rates increasing with experience and training. However, even with self-selection there is a limit to the effectiveness of these wage differentials for eliminating turnover. If the time period in which a worker is "inexperienced" is relatively short, then it may be difficult to design a wage schedule that completely compensates for the cost differences. At the limit, if training is instantaneous upon the beginning of employment, then it is impossible for the firm to pay a wage differential to "experienced" workers, for a worker becomes experienced at the very moment he is employed. In this case the only device a firm could employ is an application fee. However, its effectiveness is also quite limited. There is a moral hazard problem in that workers may foresee the firm entering the "application business" of simply collecting fees. Furthermore, workers may not have access to the capital market to borrow a possibly very large application fee.

It is surely unreasonable to explain unemployment solely on the basis of lack of knowledge of firms' characteristics. Product demand uncertainty and its role in layoffs seem to have more empirical significance. The appeal of this model rests not on its empirical validity, but on the logical structure of the analysis, and its focus on the interaction of the unemployment pool with the markets for experienced and inexperienced workers, through the costs of turnover to individual firms. While the exact formal basis for the quit decision is artificial, it does allow for a concentration on these complicated interactions without the additional complexity of an explicit model of demand uncertainty, complete with the necessity of modelling layoffs, implicit employment contracts, inventories, and other variables that would be required by a rigorous general equilibrium model.[2]

The same comment is required for the assumption that firms are unable to regulate the flow of excess applicants through a set of application fees or seniority wages. If moral hazard problems are ignored or eliminated through explicit contracts, the necessary set of markets will be complete and no involuntary unemployment will obtain in the model. On the other hand, as a practical matter, it is impossible for firms to contract away all the randomness and heterogeneity it faces in the labor market. Workers differ with respect to productivity, probability of absenteeism and quitting, and other variables that are crucial to determining a worker's value, yet are difficult to observe and write contracts on. Each of these variables could lead to incompleteness in the set of market-clearing prices required for full-employment equilibrium.

Any incompleteness in the number of prices and any uncertainty that affects the

[2]See Costas Azariadis and Martin Baily.

quit rate will enter the unemployment flows and equilibrium configuration in a manner similar to the example explored here. Thus it is useful to treat the assumptions as loose characterizations of important labor market phenomena, and build more realistic models once the logic is fully understood in a simple context.

A. The Firm's Problem

Assumption 1: *Firms produce output with employed labor E according to a nonincreasing returns production function*

$$Q = f(E), f' > 0, f'' \leqq 0$$

Assumption 2: *The capital market is ignored. However, there is a fixed cost $F \geq 0$ for setting up a firm.*

Assumption 3: *New workers (N) must be trained at the outset of employment.* Training costs (T) take place at increasing marginal costs in output terms according to

$$T = T(N) T' > 0, T'' > 0$$

Assumption 4: *Every firm is characterized by a given set of nonpecuniary job attributes.* Workers differ in preferences for these attributes. The attributes are not known to the workers upon becoming employed, but instead, they are learned by working at a firm. Once a worker learns a firm's attributes, he trades off his current wage plus nonpecuniary benefits against the expected benefit of quitting to look for another job and makes a quit decision.[3] If we let z denote a measure of labor market tightness, say the average wage rate adjusted for the probability of getting a job (and including the average nonpecuniary utility), then a firm's quit rate (q) depends on its wage w relative to z:

$$q = q(w/z), q' < 0\ q'' > 0$$

Thus, dissatisfied workers are more likely to quit the tighter are conditions in the labor market. In a stationary state, new hires equal quits.

$$N = q(w/z)E$$

Assumption 5: *The firm may hire new workers N only as long as it has enough willing applicants at its going wage rate.* The applicant function also depends on the firm's relative wage rate w/z, or

$$N \leq A(w/z), A' > 0$$

Assumption 6: *Firms are unable to charge an application fee.* This is a crucial assumption; the lack of competitive application fees is responsible for the incompleteness of markets and for the equilibrium unemployment.

Assumption 7: *The firm faces a perfectly competitive output market at a price equal to one (the numeraire) and chooses a wage of w.* Its optimization problem may be written as follows.

$$\max_{w,E,N} R = f(E) - wE - T(N) - F$$

$$\text{subject to: } N = q(w/z)E : \lambda$$
$$N \leq A(w/z) \quad : \mu$$

Letting λ and μ denote the multipliers we have the Lagrangian

$$L = f(E) - wE - T(N) - F + \lambda[N - q(w/z)E] + \mu[A(w/z) - N]$$

The first-order conditions expressing the firm's wage, employment, and new-hire tradeoffs at an interior solution $(E, w, N) > 0$ are written as follows:

$$(1) \quad E > 0, \quad f'(E) - w - \lambda q(w/z) = 0$$

$$(2) \quad w > 0, \quad -E[1 + \frac{\lambda}{z} q'(w/z)] + \frac{\mu}{z} A'(w/z) = 0$$

$$(3) \quad N > 0, \quad -T'(N) + \lambda = 0$$

In addition, we have the first-order conditions on the constraints,

$$(4) \quad \lambda[N - q(w/z)E] = 0$$

$$(5) \quad \mu[A(w/z) - N] = 0$$

[3]Alternatively, we could generate this quit-rate function if workers have a preference for job variety. For simplicity, on-the-job search is not permitted.

In order to focus on the possibility of involuntary unemployment, it is *assumed* the firm has excess applicants. From (5), we have

$$A(w/z) > N \rightarrow \mu = 0 \tag{6}$$

The remaining first-order conditions exhibit the tradeoffs facing the firm. Substituting λ from (3) into (2), we have

$$E + \frac{T'(N)}{z} q'(w/z)E = 0 \tag{7}$$

This is the wage-turnover cost tradeoff. If the firm raises its wage by a unit, direct wage costs per employee rise by E units; turnover falls by $(1/z)q'E$ units and these workers must be replaced, each at cost T'. Rewriting (7) and (3), we have

$$T'(N) = -\frac{z}{q'(w/z)} = \lambda \tag{8}$$

Substituting (3) into (1), we have

$$f'(E) = w + q(w/z)T'(N) \tag{9}$$

The marginal revenue product of an additional worker equals the marginal cost of an additional worker—the wage plus the portion of the worker's turnover costs amortized for a single period.[4]

Substituting (8) into (9), we have

$$f'(E) = w\left[1 - \frac{q(w/z)}{(w/z)q'(w/z)}\right] \tag{10}$$

Denoting the quit-rate elasticity by $\epsilon > 0$, we have a variant of the conventional monopsony formula,

$$f'(E) = w[1 + 1/\epsilon] \tag{11}$$

Noting that hires equal quits, we rewrite the constraint,

$$N = q(w/z)E \tag{12}$$

Equations (9), (10), and (12) may be solved for (E, w, N) as functions of the single exogenous parameters z. It is easy to show that[5]

$$E = E(z), \qquad E' < 0 \tag{13}$$

$$w/z = W(z), \qquad W' < 0 \tag{14}$$

$$N = N(z), \qquad N' \gtrless 0 \tag{15}$$

If $f'(E) = 0$ (constant returns to scale), then $N'(z) < 0$. In order to demonstrate the involuntary unemployment result with as little complexity as possible, we make this assumption. As the labor market tightens (z rises), the firm finds its quit rate rising. It economizes on turnover costs by lowering employment (and new hires). However, it allows its relative wage w/z to fall, implying a higher quit rate at the new optimum. Thus, the firm adjusts its wage rate to the state of the labor market, but, as we shall show in Section II, this wage flexibility is not sufficient to completely eliminate unemployment in equilibrium.

B. *Incomplete Markets, Application Fees, and Market Clearing*

The insufficiency of wage flexibility in clearing the market is a consequence of the manner in which the applicant function enters the firm's optimization. The firm faces two interrelated labor markets, an *internal* labor market for experienced (trained) employees and an *external* market for new applicants. Since the firm has only a single wage rate with which to economize on labor simultaneously in both markets, this single wage is generally unable to clear both markets simultaneously.

Because of turnover costs, the internal labor market dominates the firm's decision making in a loose (low z) market; that is, the applicant function enters merely as a nonbinding constraint (equation (6)). The possibility of a binding constraint is discussed in Section IV.

On the other hand, as David Soskice points out, the firm could economize on applicants separately by charging an application fee in order to equate applicants to new hires. Letting the fee be denoted by $\hat{a}$, we have

$$A(w - r\hat{a}/z) = N(z) \tag{16}$$

where $r\hat{a}$ is the (implicit) interest on the fee.

[4] If the quit rate is q per period, then a worker's expected tenure is $1/q$ periods. The $T'(N)$ is spread over the entire period equally. The discount rate has been set equal to zero for simplicity.

[5] See the author (1973b) for the details of these derivations.

Clearly, there exists a fee $\hat{a}$ that would eliminate the excess applicants. Furthermore, if all firms charged excess applicant fees, these fees would lower the expected returns from quitting (z) and imply a labor market equilibrium with zero structural unemployment.

Unfortunately, the use of such application fees is generally limited. Union regulations, antidiscrimination laws, and morale problems generally require firms to maintain equal pay for equal work. In addition, there is a serious moral hazard problem. If the equilibrium fee is very large, workers might (correctly) fear that a firm has entered the "application" industry; that is, it would be in the firm's interest to falsely advertise vacancies to collect application fees.

Another possibility to ensure market clearing is a rising wage structure. This policy is not considered in the optimization written previously because training takes place instantaneously.[6] If training takes time, however, then the application fee may be interpreted as the wage differential between trained and untrained workers. As before, however, this policy has only limited scope. The entire training costs must be captured during the apprenticeship program. This is impossible if training costs are so large to require a negative apprenticeship wage.[7] Furthermore, since training here is firm specific, workers may be averse to bearing such costs in the absence of explicit contractual obligations on the part of the firm.

II. Market Equilibrium

In the absence of application fees or other contractual arrangements, we may solve for the free entry equilibrium in the labor market. Formally, an equilibrium is a number of firms n and wage rates, employments, new hires, and applicants $[w_i, E_i, N_i, A_i]$ for the n firms, such that the n internal labor markets for experienced workers (quits) and n external labor markets for new applicants are cleared. Since there are only n prices attempting to clear $2n$ markets, it is not surprising that quantity rationing must serve as the clearing device in some markets, leading to the possibility of unemployment at the equilibrium.

Equations (13)–(15) summarize the demands of a single firm in this economy as a function of the aggregate variable z. For simplicity, assume that every firm has identical technology and that workers' preferences over nonpecuniary characteristics of firms are symmetric across the attributes offered.[8] Hence, no equalizing wage differentials are necessary; every firm has an identical quit-rate function and all choose identical $[w, E, N, A]$. Under these assumptions, we may easily solve for the equilibrium z^* for n, the number of firms in the market.

Let z, the summary measure of labor market tightness, equal the expected wage in the market.[9] Letting π denote the probability that an unemployed (searching) worker obtains an offer, since every firm pays an identical wage, we have

$$z = \pi w \tag{17}$$

If the equilibrium $\pi < 1$, then involuntary unemployment is positive, whereas full employment entails $\pi = 1$. The supply of workers to the market (each supplying one unit of labor) depends on the probability of employment as well as the wage. Let supply S be given by

$$S = S(z), S' > 0 \tag{18}$$

Since $nE(z)$ workers are employed and $S(z)$ workers each supply a unit of labor, the stock of involuntarily unemployed $U(z)$ is given by

$$U(z) = S(z) - nE(z) \tag{19}$$

[6]Since all trained workers are perfect substitutes, they ought to be paid identical wages at the optimum See Joseph Stiglitz.

[7]Joanne Salop and the author and A. Weiss explore models of self-selection and apprenticeship.

[8]That is, firms are equidistantly spaced in attribute space relative to preferences. For the details, see the author (1978).

[9]To be fully rigorous, z ought to denote the expected wealth stream accruing to the worker if he quits. This approximation is used for expositional convenience and does not alter the logic of the result. See the author (1973a) for the rigorous formulation.

Note that $U(z)$ does not include those workers who are frictionally unemployed. This can be illustrated by examining the functioning of the market. The state of the market before the period begins can be described as follows: Of the $S(z)$ workers in the market, $nE(z)$ are employed and $U(z)$ are unemployed. At the beginning of a period, some workers quit (a total of $Q(z) = nq(W(z))E(z)$) and enter the unemployment pool. (On-the-job search has been ignored for simplicity; it could be added without changing the basic results of the model, if it is more efficient to search while unemployed.) Thus the total number of workers searching for a job are those that were previously unemployed ($U(z)$) plus those that have just quit ($Q(z)$) or a total of $U(z) + Q(z)$. Of these workers, $nN(z)$ are hired; this is the measure of frictional unemployment in the market. If hiring is done randomly among all the applicants, the probability π that any particular searcher is hired is given by

$$(20) \qquad \pi \equiv \pi^u(z) = \frac{nN(z)}{U(z) + Q(z)}$$

Since the market is in equilibrium, hires equal quits, or

$$(21) \qquad Q(z) = nN(z)$$

Thus, $U(z)$ measures involuntary unemployment and $Q(z)$ measures frictional unemployment.

Substituting (21) and (19) into (20), we have $\pi^u(z)$ as pictured in Figure 1.

$$(22) \qquad \pi^u(z) = \frac{nN(z)}{S(z) - nE(z) + nN(z)}$$

Differentiating, we have[10] $d\pi^u/dz < 0$. Substituting $W(z)$ from (14) into the definition of z in (17), we have a second expression for π.

$$(23) \qquad \pi = \pi^w(z) \equiv \frac{1}{W(z)}$$

Differentiating, we have

$$\frac{\partial \pi^w}{\partial z} > 0, \qquad \text{since } W' < 0$$

[10]Assuming $N'(z) < 0$. Recall that constant returns production is sufficient for $N' < 0$.

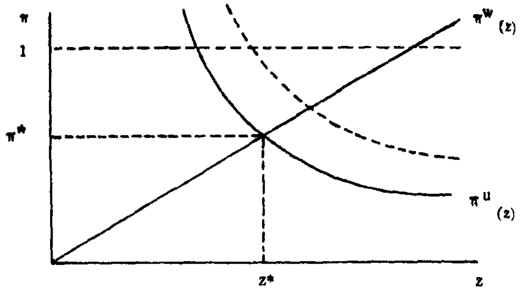


FIGURE 1. EQUILIBRIUM

We set $\pi^u(z) = \pi^w(z)$ to solve for the equilibrium value z^*, as a function of the parameter n.

$$(24) \qquad \frac{1}{W(z^*)} = \frac{nN(z^*)}{S(z^*) - nE(z^*) + nN(z^*)}$$

which defines

$$(25) \qquad z^* = z(n)$$

If $N'(z) < 0$, a unique underemployment equilibrium obtains as pictured below. $\pi^* \epsilon (0, 1)$ is also necessary.

$$(26) \qquad 0 < \pi^* < 1 \leftrightarrow U(z^*) > 0 \leftrightarrow S(z^*) > nE(z^*)$$

This depends quite crucially on the number of firms, n, as well as the supply of labor function.

As the number of firms increases, the $\pi^u(z)$ function shifts up[11] and π^* and z^* rise. Thus, we must ask the question of whether entry by new firms will continue to tighten the labor market until $\pi^* = 1$. In general, this need not be true. Suppose free entry continues until profits per firm equal zero; then rewriting profits $R(z)$ as a function of z, we have

$$(27) \qquad R(z) = f(E(z)) - zW(z)E(z) - T(N(z)) - F$$

The number of firms depends crucially on the level of fixed costs F. By setting F we can essentially set the number of firms n at any level desired. Formally, we have

$$(28) \qquad R(z) = 0$$

From (25) we have $z^* = z(n)$, $z' > 0$. Equa-

[11]Since $n = qE < E$.

tions (25) and (28) may be solved for the unique equilibrium values (z, n). Uniqueness may be demonstrated using the envelope theorem.[12]

$$\frac{dR}{dn} = \left[-wE + \frac{\partial R}{\partial w}\frac{\partial w}{\partial z} + \frac{\partial R}{\partial N}\frac{\partial N}{\partial z} + \frac{\partial R}{\partial E}\frac{\partial E}{\partial z}\right]\frac{dz}{dn}$$

Since $\partial R/\partial w = \partial R/\partial N = \partial R/\partial E = 0$ at the optimum for each firm, we have

$$\frac{dR}{dn} = -wE\left(\frac{dz}{dn}\right) < 0$$

At a zero profit level, a new entrant will incur negative profits. Thus, *an equilibrium in the labor market may exist in which the equilibrium probability of employment* (z) *is less than one.* Referring back to (26), this implies that unemployment $U(z)$ is positive.

Suppose the supply of labor function shifts, due to governmental manpower programs or migration by new workers into the economy. This rise in the $S(z)$ function lowers π in the short run as more applicants compete for the available vacancies in the market. Quits fall as employed workers perceive the worsened opportunities from search which in turn allows firms to lower wage rates. Profits rise and induce entry by new firms. Surprisingly, the new equilibrium entails an identical z as originally. We may prove this as follows.

Letting the supply shift parameter be denoted by α and rewriting the equilibrium condition (24) and free entry condition (27), we have

$$(29) \quad \pi(z) \equiv \frac{1}{W(z)} = \frac{nN(z)}{S(z,\alpha) - nE(z) + nN(z)}$$

$$(30) \quad R(z) = f[E(z)] - zw(z)E(z) - T(N(z)) - F = 0$$

Equation (30) may be solved for a unique level $\hat{z}$ for all α and n; as α changes, the equilibrium number of firms n simply adjusts to maintain equality in (30). Thus policies that increase the supply of labor to the market have no effect on the *expected* real wage in equilibrium. The proportion of these new workers who become employed is identical to the proportion previously employed.

This unemployment $U(\hat{z})$ is a permanent state of the market. Macro-economic stabilization policies cannot eliminate it. Instead, it arises from the structure of the economy—the lack of market clearing in external labor markets in conjunction with firms' monopsony power in internal labor markets. It is involuntary in the sense that the unemployed workers would be willing to accept a job at the going wage rate; however, at the going wage, offers are not forthcoming to all the unemployed. I call this unemployment *involuntary structural unemployment.* This involuntary structural unemployment is in addition to *frictional unemployment* resulting from workers quitting one job to look for another. Frictional unemployment is measured simply by new hires (or quits) of $\hat{n}N(\hat{z})$.

III. Wage Differentials, Search Unemployment, and Disguised Unemployment

The equilibrium constructed has no wage differentials. However, if firms differ in turnover costs, they will make different optimal wage-turnover tradeoffs. This is expressed in equation (8), which may be rewritten as

$$T'(N) = -\frac{z}{q'(w/z)}$$

If there are turnover-cost induced wage differentials,[13] the optimal behavior by applicants will lead to the existence of equilibrium *search unemployment.* We may model this formally as follows.

Applicants choose a firm (a queue) in order to maximize expected return. If firm j pays a wage w_j, has vacancies N_j and applicants $A_j > N_j$, the expected wage to an applicant from waiting in firm j's queue is

[12]The condition that n must equal an integer is ignored This is not an unreasonable approximation if n is fairly large.

[13]Permanent noncompensating wage differentials may also be due to differences in production functions, discount rates, etc. See the author (1973a).

given by z_j, where[14]

(31) $z_j = w_j N_j / A_j \quad j = 1, 2, \ldots, n$

Suppose there are $\bar{A}$ total applicants in the market. If each applicant observes $\{w_j, N_j, A_j\}$ and chooses a queue to max z_j, the number of applicants will adjust until an equilibrium queue distribution is achieved in which returns are identical in each, or

(32) $z_j = \bar{z}$ for all j

Solving (31) and (32) we have

(33) $A_j = N_j(w_j/\bar{z})$

(34) $\bar{A} = \sum A_j$

Clearly $\bar{z}$ will depend on $\bar{A}$ and $\{w_j, N_j\}$.

For example, suppose firms' wages were distributed uniformly in (w_a, w_b) and due to both production function and training function differentials, every firm had an identical number of vacancies N. Then solving explicitly, we have

(35) $A(w) = (w/\bar{z}) \cdot N$

This is a linear function of w. Since

(36) $\bar{A} = \int_{w_a}^{w_b} A(w)dw$

we have

(37) $\bar{z} = \left(\frac{w_b^2 - w_a^2}{2}\right) \frac{N}{\bar{A}}$

In Figure 2, the area between $A(w)$ and N consists of search unemployment[15] plus structural involuntary unemployment. On the diagram, this is shown as follows. The area between $S(w)$ and N measures search unemployment and the area between $A(w)$ and $S(w)$ measures structural involuntary unemployment. Frictional unemployment is measured as the area under N, the total flow of vacancies in the market.

It may be noted that equilibrium may entail zero involuntary unemployment. (For example, $S(w)$ could measure the *total* applicants per firm.) However, equilibrium does imply that only the minimum wage firm may have a binding queue. In equilibrium any firm choosing $w > w_a$ will have excess applicants.[16]

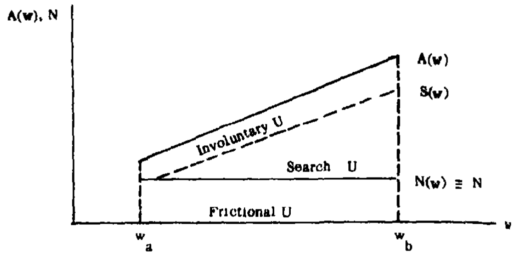


FIGURE 2. EQUILIBRIUM WITH WAGE DIFFERENTIALS

Finally, we may also measure the *disguised unemployment* that arises as a result of the existence of *structural unemployment*. If there were no structural unemployment, there would be a supply $S(w)$. However, because of the limited opportunities in the market, only $S(z)$ enter the labor force. Thus a stock of potential workers $D(z)$ never enter, where $D(z) = S(w) - S(z) \geq 0$ These workers comprise *disguised unemployment*

IV. Full-Employment Equilibrium

The solution of the formal model demonstrates only the possibility of an equilibrium with structural unemployment, not its necessity. In the analysis it is *assumed* that the necessary condition $\pi < 1$ is fulfilled. Fortunately, some supply function $S(z)$ or fixed cost F can always be found that ensures that z equilibrates at $\pi < 1$. On the other hand, for small $S(z)$, an equilibrium with $\pi = 1$ obtains, a full-employment equilibrium.

Moreover, the analysis of Section I assumes that the applicant constraint is not binding. In my 1973b paper, the possibility of a binding applicant constraint is considered, the regions where it is binding are derived, and the expanded $W(z)$, $E(z)$, and $Q(z)$ functions are calculated. Employing that expanded analysis in the present equilibrium model, involuntary structural unemployment obtains for certain values of the technological and supply parameters of

[14]As before, the expected *wealth* in each queue should be calculated.

[15]See Robert Hall for an application of this analysis.

[16]This flows directly from (38) and (39). If for the $w > w_a$ firm, $z(w) = z_a$ and $w > w_a$, then $\pi(w) = N(w)/A(w) < 1$.

the model. Moreover, cases may exist in which there are multiple equilibria, some with full employment and some with unemployment. This is no surprise, for multiple equilibria and nonexistence often occur in models of price-setting agents and incomplete markets.[17]

When the possibility of wage differentials is included as in Section III, a similar expansion of the analysis is necessary; the involuntary unemployment area between $A(w)$ and $S(w)$ may disappear. However, as long as there are wage differentials, equilibrium must entail a positive level of search unemployment, as more applicants queue at high wage than low wage firms.

V. Conclusions

An incomplete set of market-clearing wages will prevent the labor market from attaining the classical zero involuntary unemployment equilibrium. Instead a permanent level of involuntary structural unemployment and disguised unemployment may result as quantities adjust to the non-market-clearing wages. This unemployment is in addition to the frictional and search unemployment of the "new" macroeconomics.

The job shortage interacts with firms' monopsony power in the labor market to ensure that the aggregate unemployment rate is not optimal. Even if the level of frictional unemployment were efficient, the three other types of unemployment are not. Search unemployment requires an equalization of *average* rather than marginal rates of substitution; disguised and structural unemployment entail quantity rather than price adjustments.

Finally, the analysis of the paper has focused on the existence of a structural unemployment equilibrium. It should be noted that an equilibrium may also exist with zero structural unemployment. Such multiple equilibria generally exist in economies with incomplete markets or market power. However, a detailed analysis of the exact conditions under which multiple equilibria obtain is left to a sequel.

[17]See John Roberts and Hugo Sonnenschein, and the author (1978) for examples.

REFERENCES

C. Azariadis, "Implicit Contracts and Underemployment Equilibria," *J. Polit. Econ.*, Dec. 1975, *83*, 1183–202.

M. Baily, "Wages and Employment under Uncertain Demand," *Rev. Econ. Stud.*, Jan. 1974, *41*, 37–50.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

R. E. Hall, "Why is the Unemployment Rate So High at Full Employment?," *Brookings Papers*, Washington 1970, *3*, 369–402.

D. Mortensen, "A Theory of Wage and Employment Dynamics," in Edmund Phelps et al., eds., *The Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

Edmund Phelps et al., (1970a) *The Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

———, (1970b) "Money Wage Dynamics and Labor Market Equilibrium," in his *The Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

J. Roberts and H. Sonnenschein (1977), "On the Foundations of the Theory of Monopolistic Competition," *Econometrica*, Jan. 1977, *45*, 101–14.

S. C. Salop, (1973a) "Systematic Job Search and Unemployment," *Rev. Econ. Stud.*, Apr. 1973, *40*, 191–201.

———, (1973b) "Wage Differentials in a Dynamic Theory of the Firm," *J. Econ. Theory*, Aug. 1973, *6*, 321–44.

———, "Monopolistic Competition with Outside Goods," unpublished paper, Univ. Pennsylvania 1978.

J. K. Salop and S. C. Salop, "Self-Selection and Turnover in the Labor Market," *Quart. J. Econ.*, Nov. 1976, *90*, 619–27.

D. Soskice, "Salop and Stiglitz on Involuntary Unemployment," unpublished paper, Univ. California-Berkeley 1974.

J. E. Stiglitz, "Equilibrium Wage Distributions," unpublished paper, Stanford Univ. 1974.

A. Weiss, "Education as a Test," unpublished paper, Bell Laboratories 1977.

On the "Importance" of Productivity Change

By CHARLES R. HULTEN*

Robert Solow's paper on technical change provides an economic rationale for the so-called total factor productivity "residual"—the growth rate of real product not explained by the share-weighted growth rates of the real factor inputs. Solow demonstrated that under the assumptions of a Hicks-neutral aggregate production function and competitive equilibrium, the residual is equivalent to the growth rate of the Hicksian efficiency parameter, which in turn is equivalent to the rate at which the aggregate production function is shifting over time. An important implication of this result is that, under the appropriate assumptions, the shift in the production function can be measured using price and quantity data alone, without the need of estimating or assuming the values of such parameters as the elasticity of substitution between capital and labor.[1]

Although the residual is a valid measure of the shift in technology, it does not indicate the true importance of productivity change as a source of economic growth. An increase in total factor productivity will in general lead to an increase in output (as the inputs are used more efficiently) and thus to additional saving and capital formation. Part of the historically observed growth rate of capital stock is, therefore, the *result* of productivity change, and must be recognized as such when assessing the importance of productivity change as a source of growth.

This paper suggests an accounting framework for measuring the *importance* of productivity change using price and quantities alone. It is based on an intertemporal specification of technology closely related to the framework proposed by Edmond Malinvaud (1953, 1961). An effective rate of productivity change (termed the dynamic residual) is defined to be the residual growth in total consumption not explained by the rate of change of total primary input. This residual is then related to the change in the Malinvaud intertemporal production possibility frontier due to changes in total factor efficiency. Since capital accumulation is endogenous in the intertemporal framework, the dynamic residual measures the impact of annual changes in factor efficiency inclusive of the induced accumulation of capital. It thus provides a measure of the importance of productivity change in economic growth.

John R. Hicks and T. K. Rymes have also emphasized the need to measure technical change in a dynamic (capital endogenous) framework,[2] but, have implicitly (and explicitly, in the case of Rymes) rejected the conventional residual as a measure of changing technical efficiency. The main result of this paper is, however, that the conventional and dynamic residuals are complements rather than substitutes. They measure different aspects of the same process within a common analytical framework. As will be seen in Table 1, the conventional (atemporal) accounting framework is embedded in a more general

*The Urban Institute. I acknowledge the financial support of the National Science foundation in the preparation of this paper. I would also like to thank Larry Epstein, Melvyn Fuss, Dale Jorgenson, and Mieko Nishimizu.

[1]The recent empirical literature on *U.S.* productivity change includes Laurits Christensen and Dale Jorgenson (1969, 1970), Edward Denison (1962, 1967, 1974), Jorgenson and Zvi Griliches (1967), John Kendrick (1961, 1973), and Spencer Star. Estimates of the residual have varied greatly· for example, Jorgenson and Griliches (1967) obtain an average annual estimate of 0.10 percent for the period 1945–65, while Kendrick (1973) obtains 2.0 percent for the same period. For a discussion of some of the issues in productivity analysis, see the 1972 *Survey of Current Business* exchange between Denison and Jorgenson and Griliches.

[2]Richard Nelson also notes the interaction between productivity change and capital accumulation, but concentrates on the embodied technical change aspects of the problem. See also Denison (1974, pp. 133–35).

TABLE 1—INTERTEMPORAL ACCOUNTING SYSTEM

		Year	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Delivery of Investment to Year t (Intermediate Demand)				Final Demand		Value of Product
			1	2	3	4	Export of Capital	Consumption	Sum of (1) through (6)
1	Delivery of Investment From Year t	1		$R_2 I_1$	$R_3(1-\delta)I_1$	$R_4(1-\delta)^2 I_1$	$p_4(1-\delta)^3 I_1$	$p_1 C_1$	$p_1 Q_1$
2		2	-	-	$R_3 I_2$	$R_4(1-\delta)I_2$	$p_4(1-\delta)^2 I_2$	$p_2 C_2$	$p_2 Q_2$
3		3	-	-	-	$R_4 I_3$	$p_4(1-\delta)I_3$	$p_3 C_3$	$p_3 Q_3$
4		4	-	-	-	-	$p_4 I_4$	$p_4 C_4$	$p_4 Q_4$
5	Import of Capital		$R_1 K_0$	$R_2(1-\delta)K_0$	$R_3(1-\delta)^2 K_0$	$R_4(1-\delta)^3 K_0$			
6	Total Capital Outlay (Rows 1 5)		$R_1 K_0$	$R_2 K_1$	$R_3 K_2$	$R_4 K_3$			
7	Total Labor Outlay		$w_1 L_1$	$w_2 L_2$	$w_3 L_3$	$w_4 L_4$			
8	Total Outlay (Rows 6 7)		$p_1 Q_1$	$p_2 Q_2$	$p_3 Q_3$	$p_4 Q_4$			

intertemporal framework which provides the basis for computing the dynamic residual.

The other main result of this paper relates the dynamic residual to the aggregation procedure proposed by Evsey Domar for aggregating across sectors in an atemporal model. It is shown that the dynamic residual is the weighted sum of the conventional Solow residuals. This result implies that productivity analysts can obtain an empirical measure of the importance of productivity change by the appropriate averaging of the conventional residuals, using a weighting scheme based on prices and quantities used in the calculation of these residuals.

The paper has the following organization. In Section I, an intertemporal accounting framework is developed in which capital is treated as an intermediate product. In Section II, the intertemporal accounting identities are used to define both the conventional and dynamic residuals. The dynamic residual is then related to the Malinvaud intertemporal production possibility frontier and to the aggregation procedure proposed by Domar. Both residuals are then calculated in Section III using a modified version of the Christensen-Jorgenson (1969, 1970) data for the *U.S.* private domestic economy. For the period 1948–66, the conventional residual is found to average 1.55 percent per year, while the dynamic residual averages 2.76 percent. The former is found to account for 38 percent of the growth rate of real product, while the latter accounts for 66 percent and thus assigns productivity change a much larger role in the expansion of the postwar *U.S.* economy.[3]

Finally, Section IV compares the results of this paper with those of my 1975 paper. The latter are based on neoclassical growth theory, and turn out to be almost identical to the estimates of this paper. The fact that the same results are obtained under differing assumptions provides additional support for the hypothesis that productivity change was the predominant source of recent *U.S.* economic growth.[4]

I. An Intertemporal Accounting Framework

The analysis of productivity change typically starts with a detailed description of the accounting framework used in the calculations. The appropriate accounting identities are first developed, and then differentiated to obtain the productivity residual. The residual is then interpreted in

[3]Real gross output is the measure of real product underlying the 38 percent figure, while the measure underlying the 66 percent figure is a measure of real final demand. For the period 1948–66, the former grew at an average annual rate of 4.11 percent and the latter at 4.18 percent.

[4]Some authors, Kendrick (1973) for example, calculate a sufficiently large residual that they obtain this result without the adjustment for the induced accumulation of capital. If this adjustment were made, however, productivity change would be assigned an even larger role.

terms of the aggregate production function. I shall follow the procedure in developing the intertemporal analysis, and start by describing a set of intertemporal accounts in which an investment made in one period is regarded as creating an intermediate delivery (net of physical depreciation) to each subsequent year within the accounting period. Consumption is treated as a direct delivery of final demand, and labor is regarded as the only primary input.

Table 1 sets out the basic structure of the accounts for four time periods. An investment I_1 made in period 1 is delivered to production in period 2, and, assuming a constant rate of depreciation δ, $(1-\delta)I_1$, is delivered to period 3 production and $(1-\delta)^2 I_1$ is delivered to period 4 production.[5] What remains of the original investment $(1-\delta)^3 I_1$, is treated as an export to the future and entered as a component of final demand. Similarly, I_2 is delivered from period 2 to period 3, $(1-\delta)I_2$ is delivered to period 4, $(1-\delta)^2 I_2$ is exported, and so on. Consumption in each period, C_t, is treated as a direct delivery to final demand from that period. Imports of capital from the past are treated as primary inputs and are included with value-added to determine value of total primary input.

The entries in Table 1 refer to values rather than to physical quantities (or constant dollar indexes). Investment goods delivered to production are valued using the discounted rental price of capital goods R_t, which is the price paid for using the services of an investment good for one period. Consumption and the export of capital, on the other hand, are valued in terms of the discounted final sales price p_t, since all potential services associated with a commodity are transferred when consumed or when exported. Under perfect foresight, the sales price p_t is the discounted flow of rental prices (net of depreciation) plus the end-of-period value of the commodity:

$$
\begin{aligned}
(1)\quad p_0 &= R_1 + (1-\delta)R_2 + (1-\delta)^2 R_3 \\
&\quad + (1-\delta)^3 R_4 + (1-\delta)^4 p_4 \\
p_1 &= R_2 + (1-\delta)R_3 + (1-\delta)^2 R_4 \\
&\quad + (1-\delta)^3 p_4 \\
p_2 &= R_3 + (1-\delta)R_4 + (1-\delta)^2 p_4 \\
p_3 &= R_4 + (1-\delta)p_4 \\
p_4 &= p_4
\end{aligned}
$$

The convention adopted here, is that all payments take place at the end period, and that all prices are discounted to the beginning of the first period. By normalizing on the current (spot) price of the homogenous output, the present value of one unit of output can be written

$$(2)\qquad p_t = 1/\prod_{s=1}^{t}(1+r_s) \qquad t = 1,\dots,4$$

where r_s is the rate of discount implied by the intertemporal preferences of consumers. Letting R_t' denote the current rental price of capital, i.e., $R_t' = R_t[\Pi_s(1+r_s)]$, (1) and (2) imply

$$(3)\qquad R_t' = r_t + \delta \qquad t = 1,\dots,4$$

which indicates that the current rental price of capital services equals the opportunity cost of not consuming the commodity in the current period, plus the cost of the commodity used up in production.

Using (1), it is evident that summation across the first five columns of Table 1 yields the value of gross investment: $p_1 I_1$, $p_2 I_2$, $p_3 I_3$, $p_4 I_4$, respectively, for rows (1) through (4). Addition of the consumption column (6) yields the value of product originating in each period (column (7)):

$$(4)\qquad V_t = p_t Q_t = p_t I_t + p_t C_t \qquad t = 1,\dots,4$$

where V_t denotes the value of product and Q_t the quantity of the homogenous product.

Equation (4) relates the value of product sold to the value of product purchased, and corresponds to the "uses of income" side of conventional frameworks. The "sources of income" side of the accounts can be developed using a perpetual inventory definition of capital stock. Given the constant

[5] The assumption of a constant rate of depreciation is not essential for the analysis, and is used for purposes of exposition. A more general approach would be to replace $(1-\delta)^t$ with Φ_t, the relative asset efficiency. See Jorgenson (1973) for a detailed discussion of this point and for an exposition of the asset pricing model underlying (1), (2), and (3) which emphasizes the duality between replacement and depreciation.

rate of depreciation δ, capital stock in each period is given by

$$(5)\quad \begin{aligned} K_1 &= I_1 + (1-\delta)K_0 \\ K_2 &= I_2 + (1-\delta)I_1 + (1-\delta)^2 K_0 \\ K_3 &= I_3 + (1-\delta)I_2 + (1-\delta)^2 I_1 \\ &\qquad + (1-\delta)^3 K_0 \\ K_4 &= I_4 + (1-\delta)I_3 + (1-\delta)^2 I_2 \\ &\qquad + (1-\delta)^3 I_1 + (1-\delta)^4 K_0 \end{aligned}$$

Using (5), it is evident that summation of the first five elements in columns (1)-(4) of Table 1 yields R_1K_0, R_2K_1, R_3K_2, and R_4K_3, respectively. This is the total capital outlay for each of the four production periods. By adding total outlay for labor services, row 7, to the column totals, we arrive at total outlay, or value of product as measured from the sources of income side:

$$(6)\quad V_t = w_tL_t + R_tK_{t-1} = p_tQ_t \qquad t = 1,\ldots,4$$

This is row 8 of Table 1. Since endogenously produced capital is treated as an intermediate good, (6) is equivalent to value of product as defined in (4) and is not equal to value-added. Value of product, can, in fact, be allocated between value of primary input, U_t and value of intermediate input X_t':

$$(7)\quad V_t = U_t + X_t' \qquad t = 1,\ldots,4$$

Value of intermediate input is the sum of rows 1 through 4; that is, the sum of deliveries of investment periods 1 to 4 to period t. The import of capital from the past (row (5)) is treated as a primary input since it is not produced within the economic system. Payments to primary factor input in each period are thus the sum of value-added, w_tL_t, and value of imports; thus $U_t = w_tL_t + R_t(1-\delta)^{t-1}K_0$.

Value of product V_t can also be allocated on the uses side between deliveries to intermediate demand and deliveries to final demand. The former is the value of investment delivered from period t to production in each subsequent period. Final demand is the value of consumption at time t plus the value of the delivery of exports at the end of the period:

$$(8)\quad Y_t = p_tC_t + p_t(1-\delta)^{4-t}I_t$$

$$(9)\quad X_t = \sum_{s=1}^{4-t} R_{t+s}(1-\delta)^{s-1}I_t \qquad t = 1,\ldots,4$$

Equation (8) defines the value of final demand and corresponds to the sum of columns (5) and (6), exports plus consumption. Equation (9) defines the value of deliveries to intermediate demand, and is the sum of columns (1) through (4). Value V_t can thus be written

$$(10)\quad V_t = p_tC_t + p_tI_t = Y_t + X_t \qquad t = 1,\ldots,4$$

For the accounting period as a whole, the present value of gross product originating within the period is

$$(11)\quad \sum_t p_tQ_t = \sum_t p_tI_t + \sum_t p_tC_t$$

and the present value of factor payments "originating" is

$$(12)\quad \sum_t p_tQ_t = \sum_t w_tL_t + \sum_t R_tK_{t-1}$$

In order to arrive at total gross product, the re-export of capital must be added to gross product originating so that the total export of capital p_NK_N is fully accounted for. Denoting re-export by $M_4 = p_4(1-\delta)^4K_0$, total gross product is $\Sigma p_tQ_t + M_4 = \Sigma Y_t + \Sigma X_t + M_4$, where

$$\begin{aligned} \sum X_t &= \sum p_tI_t - (p_4K_4 - M_4) = \\ \sum X_t' &= \sum R_tK_{t-1} - (p_0K_0 - M_4) \end{aligned}$$

and

$$\begin{aligned} \sum Y_t &= \sum p_tC_t + (p_4K_4 - M_4) = \\ \sum U_t &= \sum w_tL_t + (p_0K_0 - M_4) \end{aligned}$$

Together, the aggregate identities imply the equality between the net addition to wealth during the accounting period and total value-added:

$$(13)\quad \sum_t p_tC_t + p_4K_4 - p_0K_0 = \sum_t w_tL_t = W_4$$

Equation (13) is the basic accounting identity used in the intertemporal productivity analysis. It is the intertemporal analogue of the conventional identity between gross national product from the expenditure side of the accounts and gross national product from the income side. The intertemporal

accounting framework can be extended to N time periods in a straightforward way.

II. A Model of Productivity Change with Endogenous Capital Stock

The conventional residual is defined as the growth rate of real product not explained by the share-weighted growth rates of the real factor inputs. The residual can be derived from equation (6) by total differentiation:

$$(14) \quad \frac{Dp_t}{p_t} + \frac{DQ_t}{Q_t} = \frac{w_t L_t}{V_t}\left[\frac{Dw_t}{w_t} + \frac{DL_t}{L_t}\right] + \frac{R_t K_{t-1}}{V_t}\left[\frac{DR_t}{R_t} + \frac{DK_{t-1}}{K_{t-1}}\right] \quad t = 1, \ldots, N$$

Rates of change, for example, DL_t/L_t, may then be replaced by actual period-by-period growth rates to produce the discrete time approximation to time derivations. In practical applications, the growth rates are approximated by logarithmic differences, for example, $ln L_t - ln L_{t-1}$, and the weights are averaged between periods, for example, by

$$\frac{1}{2}\left[\frac{w_t L_t}{V_t} + \frac{w_{t-1} L_{t-1}}{V_{t-1}}\right]$$

The result is the Törnqvist-Theil approximation to the Divisia index (see W. E. Diewert).

The separation of price and quantity components in (14) yields a discrete time version of the conventional residual:[6]

$$(15) \quad T_t^* = \frac{DQ_t}{Q_t} - \frac{w_t L_t}{p_t Q_t}\frac{DL_t}{L_t} - \frac{R_t K_{t-1}}{p_t Q_t}\frac{DK_{t-1}}{K_{t-1}} \quad t = 1, \ldots, N$$

As shown by Solow, equation (15) can be related to the underlying technology by assuming (a) the existence of a Hicks-neutral constant returns production function for each year:

$$(16) \quad Q_t = A_t F(K_{t-1}, L_t) \quad t = 1, \ldots, N$$

where A_t is the Hicksian index of productive efficiency, and (b) that the factors of production are paid the value of their marginal product:

$$(17) \quad \partial Q_t/\partial L_t = w_t/p_t; \qquad \partial Q_t/\partial K_{t-1} = R_t/p_t$$

Total logarithmic differentiation of (16), and use of the marginal productivity conditions, leads directly to (15). This implies that the residual is equivalent to the rate of change of the Hicksian efficiency parameter, (i.e., $T_t = DA_t/A_t$). The residual can therefore be interpreted as the shift in the aggregate production function (16), and the share-weighted growth rates of the inputs as a movement along the function.

While T_t^* is a valid measure of the rate of change of the aggregate technology, it does not measure the contribution of technology to growth since the induced expansion of capital is not taken into account. A measure which does allow for the expansion in capital can be defined within the accounting framework of the preceding section, since capital is treated as an endogenously determined flow. Equations (12) and (13) are the basic identities relating the real product (the present value of final demand) to primary inputs (the present value of real labor services). Total differentiation of these equations results in

$$(18) \quad \sum_{t=1}^{N} \frac{p_t C_t}{W_N}\left[\frac{Dp_t}{p_t} + \frac{DC_t}{C_t}\right] + \frac{p_N DK_N}{W_N}\left[\frac{Dp_N}{p_N} + \frac{DK_N}{K_N}\right] - \frac{p_0 K_0}{W_N}\left[\frac{Dp_0}{p_0} + \frac{DK_0}{K_0}\right] = \sum_{t=1}^{N} \frac{w_t L_t}{W_N}\left[\frac{Dw_t}{w_t} + \frac{DL_t}{L_t}\right]$$

where $W_N = \Sigma p_t C_t + p_N K_N - p_0 K_0$ is the wealth accumulated over the accounting period. The separation of prices and quantities, and the analogy with (15) and (16), results in

[6]An alternating approach would be to assume that the technologies (16) and (20) have the translog form developed by Christensen, Jorgenson, and Lawrence Lau, and develop the indexes of productivity change using the Törnqvist-Theil approximation. The resulting discrete indexes would then be "exact" for the translog form (see Diewert).

$$(19)\quad T(0,N) = \sum_t \frac{p_t C_t}{W_N}\frac{DC_t}{C_t} + \frac{p_N K_N}{W_N}\frac{DK_N}{K_N} - \frac{p_0 K_0}{W_N}\frac{DK_0}{K_0} - \sum_t \frac{w_t L_t}{W_N}\frac{DL_t}{L_t}$$

where $T(0, N)$ is the residual growth in wealth not accounted for by the growth of primary input for the period as a whole. Capital is determined endogenously and appears as either consumption or as the export (or bequest) of capital carried forward to the future. When actual growth rates are substituted for the differentials DC_t/C_t, etc., the residual $T(0, N)$—the dynamic residual—is an approximate measure of primary factor productivity which takes into account the response of capital accumulation to the change in factor efficiency.

The dynamic residual has a rather straightforward interpretation in terms of the Malinvaud intertemporal possibility frontier. This frontier defines the efficient combinations of consumption and terminal capital obtainable from the endowments of labor and initial capital, and the given levels of factor efficiency:

$$(20)\quad \Phi(C_1,\ldots,C_N,K_N;K_0,L_1,\ldots,L_N,A_1,\ldots,A_N) = 0$$

Given the linear homogeneity of the technologies (16), (20) implies

$$(21)\quad \sum_{t=1}^{N}\frac{\partial\Phi}{\partial C_t}C_t + \frac{\partial\Phi}{\partial K_N}K_N - \frac{\partial\Phi}{\partial K_0}K_0 = \sum_{t=1}^{N}\frac{\partial\Phi}{\partial L_t}L_t = W'_N$$

Furthermore, logarithmic differentiation of (20) yields

$$(22)\quad T'(0,N) = \sum_{t=1}^{N}\frac{(\partial\Phi/\partial A_t)A_t}{W'_N}\frac{DA_t}{A_t} = \sum_{t=1}^{N}\frac{(\partial\Phi/\partial C_t)C_t}{W'_N}\frac{DC_t}{C_t} + \frac{(\partial\Phi/\partial K_N)K_N}{W'_N}\cdot\frac{DK_N}{K_N} - \frac{(\partial\Phi/\partial K_0)K_0}{W'_N}\frac{DK_0}{K_0} - \sum_{t=1}^{N}\frac{(\partial\Phi/\partial L_t)L_t}{W'_N}\frac{DL_t}{L_t}$$

A condition of intertemporal equilibrium is that the gradient of Φ be proportional to the discounted shadow prices of capital, labor, and consumption goods (the discrete multiperiod model is just a transposition of the atemporal multisector model). Given this condition, the second equality in (22) is equivalent to (19), implying that the dynamic residual is equivalent to the change in the frontier (20) associated with the changes in the Hicksian A_t.

This result is the intertemporal analogue of Solow's justification of the conventional residual. Where the conventional residual is associated with the shift in the aggregate production function (16), the dynamic residual can be associated conceptually with the expansion of the intertemporal consumption possibility set due to technical advance. It is important to emphasize that both residuals can be calculated from price and quantity data alone. No assumption is made about the parametric form of the technology (for example, Cobb-Douglas), or about the parametric form of consumer preferences.

The dynamic residual can be shown to be a linear combination of the conventional residuals. This follows by differentiating (4) totally and substituting the result into (19). Using (15) to eliminate DQ_t/Q_t, this yields

$$(23)\quad T(0,N) = \sum_t \frac{p_t Q_t}{W_N} T_t^*$$

The dynamic residual is thus the weighted sum of the conventional residuals. Equation (23) is essentially the same aggregation formula first discussed (in the context of intermediate input) by Domar, and further analyzed in my forthcoming paper. Since the model of this paper involves a transposition of an interindustry model to an intertemporal model where capital is treated as the intermediate good, it is not surprising, in view of Malinvaud's results, that (23) should turn out to be the Domar aggregate rate of productivity change. The dynamic residual, (23), can thus be interpreted as an average rate for the accounting period as a whole. And, since it is shown in the author (forthcoming) that the Domar weighting

scheme has the effect of taking into account the expansion in intermediate input forthcoming as a result of productivity change, it follows from (23) that the dynamic residual (19) takes into account the effects of the induced expansion of the stock of capital.[7]

The practical implications for the measurement of productivity change are worth noting. Productivity analyses produce sequences of residuals $(T_1^*, \ldots, T_N^*)$; these residuals are typically averaged to obtain a summary measure for the N periods taken as a whole. The summary measure is then interpreted as the average rate of shifting of the aggregate production function. This paper suggests that another averaging procedure is also useful. When the annual residuals are averaged using the Domar weights, the result is a measure of the importance of productivity change. Both averaging procedures produce insights into the process of technical change, and it is a mistake to use the simple (or geometric) average of the T_t^* to infer both the magnitude of technical change and its role in the process of economic growth.

III. Measurement

The conventional and dynamic residuals defined in the preceding section were calculated using data for the *U.S.* private domestic economy by Christensen and Jorgenson (1969, 1970). The Christensen-Jorgenson study distinguishes seven types of capital input and adjusts labor input for hours worked and education attainment. On the product side, consumption and investment goods are distinguished and deflated separately.

The Christensen-Jorgenson data must be modified to fit the one-sector framework of this paper. A single deflator is used for both consumption and investment, and capital stock is estimated from the total investment series using a perpetual inventory method and an average rate of replacement of .0712. The discount rate used in the present value calculations is derived from the constant returns assumption and from equation (3).[8]

Table 2 presents the annual growth rate of the various residuals and variables used in the calculation, for the period 1948–66.[9] Real gross product grew at an average annual rate of 4.11 percent for this period, and the conventional residual at 1.55 percent per year. Labor's share of income is approximately 0.6, so that the share-weighted contribution of labor is 0.88 percent, and the share-weighted contribution of capital is 1.66 percent. Under the usual interpretation, productivity change thus "explains" 38 percent of the growth rate of gross product, labor 21 percent, and capital 41 percent.

The dynamic residuals are given in column (1) of Table 3. The initial year of the accounting period is in all cases 1948, and the terminal year varies from 1949 to 1966. The dynamic residual for the period 1948–66 is 2.76 percent; this explains 66 percent of the growth rate of the real addition to wealth:

$$\overline{W}_N = \sum_t (p_t C_t / W_n)(DC_t/C_t) + (p_N K_N / W_N)(DK_N/K_N) - (p_0 K_0 / W_N)(DK_0/K_0)$$

Essentially the same picture prevails from 1959 to 1966. Before 1959, the dynamic residuals are larger, and the percentage of $\overline{W}_N$ explained higher. This reflects the large conventional residuals prevailing in the early periods of the study, as well as the patterns of consumption and investment (see columns (6) and (7) of Table 2).[10]

The average annual growth rate of the

[7]The weights in (21) sum to a quantity which is greater than or equal to one, since $\Sigma_t p_t Q_t \geqq \Sigma_t w_t L_t$ by (12). The weights thus magnify the effects of the annual residuals T_t^*.

[8]The capital benchmark used in the perpetual inventory calculation is also derived from the Christensen-Jorgenson data. The rate of discount is calculated from the equation $p_t Q_t = w_t L_t + r_t K_{t-1} + \delta K_{t-1}$.

[9]The period 1948–66 was chosen in order to compare the results with those presented in my 1975 paper. The conventional and dynamic residuals were also calculated for a sample of subperiods between 1948–66.

[10]The fall in consumption in 1949 explains the peculiar result that the dynamic residual explains 114.6 percent of the growth in real final demand between 1948–49, and the very high growth rate of investment between 1948 and 1958 helps explain the large numbers in columns (1) and (2) of Table 3.

TABLE 2—PERCENTAGE ANNUAL GROWTH RATES OF SELECTED VARIABLES, 1948–66

Year	Conventional Residual (1)	Real Product (2)	Real Labor Input (3)	Real Capital Input (4)	Labor's Share of Income (5)	Real Consumption (6)	Real Investment (7)
1948	3.56	6.13	2.04	4.02	60.9	4.89	9.28
1949	0.33	0.03	−3.07	3.68	60.8	−0.03	0.81
1950	6.49	9.97	3.33	5.55	58.1	5.03	20.26
1951	1.76	6.62	4.36	5.92	57.8	5.59	8.61
1952	−0.59	2.53	1.12	5.13	58.5	4.20	−0.62
1953	2.08	5.01	1.44	5.25	59.6	4.41	6.13
1954	−1.11	−0.82	−3.23	4.03	58.5	1.41	−5.55
1955	4.17	7.90	3.52	5.14	57.8	4.99	13.94
1956	−1.13	2.33	2.29	4.75	59.0	2.61	1.77
1957	−0.37	1.55	−0.04	4.26	59.2	2.12	0.31
1958	−1.30	−1.05	−2.58	2.89	58.6	2.78	−8.92
1959	2.83	6.32	3.92	3.62	58.6	4.57	10.16
1960	0.01	2.57	1.83	3.23	59.0	3.87	−0.21
1961	0.95	1.95	−0.58	2.71	58.4	3.67	−2.10
1962	3.45	6.03	2.49	3.35	57.8	4.66	9.23
1963	1.73	4.00	1.46	3.48	57.5	3.67	4.64
1964	2.46	5.40	2.54	3.78	57.5	7.19	6.30
1965	2.61	6.25	3.53	4.12	56.6	3.76	7.10
1966	1.96	6.10	4.15	4.33	56.7	6.18	5.98
Average Annual Growth Rate	1.55	4.11	1.47	4.17		3.95	4.36

TABLE 3—THE IMPORTANCE OF TECHNICAL CHANGE UNDER CONVENTIONAL AND FISHERIAN RESIDUALS (Shown in Percent)

Years	Dynamic Residual $T(1948, t)$ (1)	Average Growth Rate of Real Product, $\overline{W}_N$ 1948-t (2)	Ratio (1)/(2) × 100 (3)	Average Growth Rate of Conventional Residual 1948-t (4)	Average Growth Rate of Real Product, Q_t 1948-t (5)	Ratio (4)/(5) × 100 (6)
1948–49	3.38	2.95	114.6	1.93	3.04	63.6
1948–50	5.99	6.76	88.6	3.43	5.30	64.9
1948–51	5.27	6.89	76.5	3.01	5.63	53.4
1948–52	4.10	5.63	72.8	2.28	5.00	45.4
1948–53	3.99	5.51	72.4	2.24	5.00	44.9
1948–54	3.25	4.18	77.8	1.76	4.15	42.4
1948–55	3.69	4.90	75.3	2.06	4.61	44.6
1948–56	3.14	4.45	70.6	1.70	4.36	40.0
1948–57	2.81	4.01	70.1	1.49	4.07	36.6
1948–58	2.46	3.39	72.6	1.23	3.60	34.3
1948–59	2.61	3.74	69.8	1.37	3.82	35.7
1948–60	2.45	3.62	67.7	1.26	3.72	33.9
1948–61	2.39	3.47	68.9	1.24	3.60	34.4
1948–62	2.57	3.72	69.1	1.38	3.76	36.8
1948–63	2.58	3.74	69.0	1.41	3.77	37.3
1948–64	2.77	3.99	69.4	1.47	3.87	37.9
1948–65	2.74	4.06	67.5	1.53	4.01	38.3
1948–66	2.76	4.18	66.0	1.55	4.11	37.8

conventional residual is included in Table 3 for purposes of comparison, and column (6) indicates the percentage of gross output explained by this residual. Although this ratio increases as the accounting period shortens, it is uniformly less than the ratio of the dynamic residual to the growth rate of final demand set out in column (3).

Tables 2 and 3 provide support for the following conclusion: when the reproducibility of capital is taken into account, the relative importance of efficiency change is greatly increased. For the accounting periods 1948–58 through 1948–66, the average relative importance of the conventional residual is 36 percent, while the average relative importance of the dynamic residual is 69 percent. In other words, when the extra capital forthcoming as a result of productivity change is taken into account, the importance of productivity change nearly doubles.

IV. Conclusion

Almost identical results were obtained in my 1975 paper using the same data base but a methodology based on neoclassical growth theory. There two models were developed: a one-sector model of optimal growth and a two-sector model of short-run equilibrium growth. In the first model, a rate of technical change was defined as the rate of shift of the aggregate production function, measured along the long-run optimal growth path. This rate termed "the long-run Fisherian rate" was then related to the Hicksian rate and bias of technical change (which are defined in terms of the shift in the production function measured along a constant capital-labor ratio.) Assuming, as before, that technical change is Hicks-neutral, the expression for the long-run Fisherian rate is given by[11]

$$(24) \qquad Z_L^* = \left[1 + \frac{\pi}{1-\pi}\sigma\right] T^*$$

where σ is the elasticity of substitution between capital and labor, π is capital's share in income, and Z_L^* denotes the long-run Fisherian rate. If σ is assumed to equal one, and π and T^* are assigned their 1948–66 average values (0.4 and 1.55 percent, respectively), Z_L^* equals 2.58 percent. The dynamic residual, $T(0, N)$, is 2.66 percent for the same period.

The two-sector model is based on Peter Diamond's article. This model is used to project the path of output $\hat{Q}^*$, and capital stock, $\hat{K}^*$, which would have occurred in the absence of technical change. In making this calculation, the elasticity of substitution between capital and labor σ is assumed to equal one in both consumption and investment sectors, and the savings rate is assumed to be constant. The effect of productivity change is then defined as the difference between the actual growth rate of output Q^*, and the hypothetical rate $\hat{Q}^*$; the implied rate (termed the short-run Fisherian rate) is given by[12]

$$(25) \quad Z_S^* = Q^* - \hat{Q}^* - \pi\hat{K}^* - (1-\pi)L^*$$

From (15), this implies

$$(26) \qquad Z_S^* = T^* + \pi(K^* - \hat{K}^*)$$

The short-run Fisherian rate is thus equal to the conventional residual plus the weighted change in capital stock forthcoming as a result of productivity change (the induced accumulation effect).

When the Christensen-Jorgenson data were used to estimate the short-run model, $\hat{K}^*$ was found to average 1.49 percent per year, and Z_S^* was found to average 2.67 percent. This leads to the conclusion that all three estimation procedures lead to approximately the same result; when the induced accumulation of capital is taken into account, productivity change explains between 2.6 and 2.8 percent of the growth rate of real product over the period 1948–66. Since the growth rate of real gross output and real final demand average 4.11 and 4.18 percent, respectively, productivity change explains around two-thirds of the economic growth under each of the three approaches.

[11]Time subscripts are dropped here (and below) for simplicity of exposition. Unless explicitly stated, all variables are assumed to depend on time.

[12]The assumption of Cobb-Douglas technologies ($\sigma = 1$) allows us to treat π as constant; this assumption is necessary for (25) and (26) to be valid.

To summarize: the long-run Fisherian rate, short-run Fisherian rate, and dynamic residual were all derived in order to estimate the *effects* of productivity change rather than its actual magnitude. The primary effect considered was the expansion in capital stock resulting from the extra saving made possible by an increase in factor efficiency. Each estimator approaches this problem with a different set of assumptions. The long-run Fisherian rate is derived from a Ramsey utility function and a constant rate of time preference,[13] the short-run Fisherian rate assumes a constant rate of saving. Both rates assume a unitary elasticity of substitution. The dynamic residual assumes that the estimated rate of return on capital reflects the rate of time preference, or, in other words, that observed prices correctly reflect intertemporal decisions. The dynamic residual unlike the first two rates can be calculated directly from price and quantity data, without having to assume specific parameter values for the savings rate of the elasticity of substitution. The dynamic residual can in fact be calculated using essentially the same sort of data used in the calculations of the conventional residual.

[13]The Ramsey utility function used in the analysis has the form

$$U = \sum_{t=0}^{\infty} a^t u(C_t/L_t)$$

where a is the constant rate of time preference.

REFERENCES

L. R. Christensen and D. W. Jorgenson, "The Measurement of U.S. Real Capital Input, 1929-1967," *Rev. Income Wealth*, Dec. 1969, *15*, 293-320.

——— **and** ———, "U.S. Real Product and Real Factor Input, 1929-1967," *Rev. Income Wealth*, Mar. 1970, *16*, 19-50.

———, ———, **and L. J. Lau,** "Transcendental Logarithmic Production Frontiers," *Rev. Econ. Statist.*, Feb. 1973, *55*, 28-45.

Edward F. Denison, "The Sources of Economic Growth in the U.S. and the Alternatives Before Us," suppl. paper no. 13, Committee for Economic Development, New York 1962.

———, *Why Growth Rates Differ: Postwar Experience in Nine Western Countries*, Washington 1967.

———, "Some Major Issues in Productivity Analysis: An Examination of the Estimates of Jorgenson and Griliches," *Surv. Curr. Bus.*, May 1969, *49*, 1-28.

———, "Final Comments," *Surv. Curr. Bus.*, May 1972, *52*, 95-110.

———, *Accounting for United States Economic Growth, 1929-1969*, Washington 1974.

P. A. Diamond, "Disembodied Technical Change in a Two-Sector Model," *Rev. Econ. Stud.*, Apr. 1965, *32*, 161-68.

W. E. Diewert, "Exact and Superlative Index Numbers," *J. Econometrics*, May 1976, *4*, 115-45.

E. Domar, "On the Measurement of Technical Change," *Econ. J.*, Dec. 1961, *71*, 710-29.

John R. Hicks, *Capital and Growth*, Oxford 1965.

C. R. Hulten, "Growth Accounting with Intermediate Inputs," *Rev. Econ. Stud.*, forthcoming.

———, "Technical Change and the Reproducibility of Capital," *Amer. Econ. Rev.*, Dec. 1975, *65*, 956-65.

D. W. Jorgenson, "The Economic Theory of Replacement and Depreciation," in Willy Sellekaerts, ed., *Essays in Honor of Jan Tinbergen*, New York 1973.

——— **and Z. Griliches,** "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 349-83.

——— **and** ———, "Issues in Growth Accounting: A Reply to Edward F. Denison," *Surv. Curr. Bus.*, May 1972, *52*, 65-94.

——— **and** ———, "Final Reply," *Surv. Curr. Bus.*, May 1972, *52*, 111.

John W. Kendrick, *Productivity Trends in the United States*, Princeton, 1961.

———, *Postwar Productivity Trends in the United States, 1948-1969*, New York 1973.

E. Malinvaud, "Capital Accumulation and Efficient Allocation of Resources," *Econometrica*, Apr. 1953, *21*, 233-69.

———, "The Analogy Between Atemporal and Intertemporal Theories of Resource Allocation," *Rev. Econ. Stud.*, June 1961, *28*, 143–60.

R. Nelson, "Aggregate Production Functions and Medium-Range Growth Projections," *Amer. Econ. Rev.*, Sept. 1964, *54*, 576–606.

T. K. Rymes, *On Concepts of Capital and Technical Change*, Cambridge 1971.

R. Solow, "Technical Change and the Aggregate Production Function," *Rev. Econ. Statist.*, Aug. 1957, *39*, 312–20.

S. Star, "Accounting for the Growth of Output," *Amer. Econ. Rev.*, Mar. 1974, *64*, 123–35.

Vertical Integration of Successive Oligopolists

By M. L. Greenhut and H. Ohta*

Vertical integration of successive monopolists (with fixed production coefficients) has long been known to provide merging monopolists with greater profit and their customers with greater outputs at lower prices. We contended in our earlier papers that similar welfare attributes apply to mergers between monopolist input suppliers and Cournot-type oligopolists.[1] But what is the result when the input supplier is also an oligopolist? The present paper answers this question. It demonstrates, in particular, that when vertical integration of successive oligopolists is mutually profitable, industry output increases and product price is lowered. The welfare gain stemming from vertical integration is further shown to hold not only under Cournot oligopoly but Stackelberg "leader-follower" type of oligopoly.

*Professor, Texas A&M University, and associate professor, Aoyama Gakuin University (visiting professor, University of Houston), respectively. We wish to thank T. Copp, S. Holmes, F. Ryan, and T. Saving for their helpful suggestions and critiques. This paper is based in part on research funded by the National Science Foundation.

[1]The possibility of perverse welfare effects resulting from vertical integration of firms subject to variable input proportions was considered initially by John Vernon and Daniel Graham, examined further by Richard Schmalensee, George Hay, Frederick Warren-Boulton, recently repeated by Martin Perry, and by John Haring and David Kaserman. However, one can hardly neglect the quantitative importance of the vast set of industries and integrations for which fixed proportions apply. For example, in the petroleum industry, the quantity of crude remains invariant from the recovery process to the shipping of crude from wells to refineries. This invariance also applies to the shipment of the refined products to dealers and their final distribution to consumers. In refining, a preselected method of converting the crude into a certain output of refined products is used. Simple substitution of another mineral for some of the crude cannot be effected without shutting down the refinery and changing the entire production process.

I. Independent Upstream-Downstream Oligopolists

Consider two vertically related activities, say in the field of energy. Let the upstream stage involve the combined operation of refining and shipping petroleum by n independent input suppliers to m independent distributor-dealers. Conceive of the product sold by the m downstream firms as involving homogeneous tankloads of gasoline. The total tankloads for sale q_L are defined to be in fixed proportion to the homogeneous inputs of gasoline, which, in turn, are received in tankwagon quantities q_w from the upstream suppliers. In effect we have

$$(1) \qquad q_L = \alpha q_w$$

where α stands for the constant coefficient of production.

Conceive next of the market demand for the final product as involving negligible cross elasticities of demand. Let this demand be a uniquely decreasing function of market price, given by

$$(2) \qquad \begin{aligned} p_L &= f(q_L) \qquad & f' < 0 \\ &= f(\alpha q_w) \qquad & \text{via (1)} \end{aligned}$$

Market demand must be equated with industry supply. This requirement establishes

$$(3) \qquad q_L = \sum_{i=1}^{m} q_{Li}$$

where q_{Li} stands for the ith distributor's supply.

Profit for the ith distributor π_i can then be given by

$$(4) \qquad \begin{aligned} \pi_i &= p_L q_{Li} - p_w q_{wi} \qquad & i = 1, 2, \ldots, m \\ &= (\alpha f - p_w) q_{wi} \qquad & \text{via (1)} \end{aligned}$$

where p_w stands for the price of the gasoline tankwagon purchased by the ith firm.[2]

[2]Other costs (such as in distribution) are disregarded in this paper without loss of generality.

Assume the m distributors are perfect competitors in purchasing gasoline. At the same instance, assume for the moment they are Cournot-type oligopolists in retailing the gasoline. The first-order profit-maximizing conditions for these distributors requires

$$(5) \qquad \partial \pi_i / \partial q_{wi} = 0$$
$$\therefore \alpha(f + f'\alpha q_{wi}) = p_w \qquad i = 1, 2, \ldots, m$$

where under perfect competition in buying, p_w is a parameter invariant to changes in the quantity purchased by the individual distributor.[3]

Equation (5) thus depicts the individual distributor's demand for gasoline in tankwagon quantities q_{wi} at market price p_w. The total market demand for gasoline tankwagons simply involves summing (5) for all $i = 1, 2, \ldots, m$.[4] This aggregation establishes

$$(6) \qquad \alpha\left(mf + \sum_{i=1}^{m} f'\alpha q_{wi}\right) = mp_w$$
$$\alpha\left(f + \frac{1}{m}\sum f'\alpha q_{wi}\right) = p_w$$

Since market demand, q_w must equate with the market supply of the n upstream firms, we also obtain

$$(7) \qquad q_w = \sum_{j=1}^{n} q_{wj}$$

where q_{wj} stands for the jth firm's supply of tankwagons of gasoline. The aggregate upstream supply depends on the profits

$$(8) \qquad \pi_j = p_w q_{wj} - c_j q_{wj} \qquad j = 1, 2, \ldots, n$$

where c_j is a constant unit (= marginal) cost of production. Without loss of generality, we are conceiving, in effect, of a two-stage industry with petroleum refining and shipment belonging to the primary (upstream) stage of production, and distributing the final product to the consumers serving as the second (downstream) stage.[5] Profit maximization by the upstream Cournot oligopolists simply requires

$$(9) \qquad \partial \pi_j / \partial q_{wj} = 0$$
$$\alpha\left(g + \frac{1}{m} f''\alpha^2 q_w q_{wj} + \frac{m+1}{m} f'\alpha q_{wj}\right) = c_j$$
$$j = 1, 2, \ldots, n$$

where $g = f + [1/m] f'\alpha q_w$, which henceforth will be called the correspondent to f. Note further that (9) is based also on profit-maximizing equilibrium conditions for the downstream producers; equation (5), in effect, is contained in (9). Summing (9) over all j's then establishes

$$(10) \qquad \alpha\left(g + \frac{1}{mn} f''\alpha^2 q_w^2 + \frac{m+1}{mn} f'\alpha q_w\right) = \frac{1}{n}\sum_{j=1}^{n} c_j$$

The equilibrium output of gasoline (in tankwagons as well as tankload quantities) is determined uniquely by (10), provided the c_j's are known and f is well-behaved.[6] Equation (10) has therefore established the equilibrium conditions for two vertically related activities, where the upstream stage involves Cournot oligopoly in selling and the downstream stage is characterized by perfect competition in buying.

II. Vertical Integration of Successive Oligopolists

The question may now be answered whether or not industry supply increases when some firms in the related stages of production integrate vertically. To answer this question, we shall consider the two

[3]The price of gasoline tankloads, on the other hand, is expected to vary as each distributor alters his supply while conjecturing that other sellers' supplies are fixed. The term $\alpha f'\alpha q_{wi}$ in equation (5) stems from this Cournot condition.

[4]Remember, in passing, that the demand for tankwagons of gasoline is readily transformed back to, in fact derives from, the downstream market demand for gasoline *tankloads*.

[5]Note that c_j could have been considered as the market price in gasoline units of the petrol-converted oil if we had assumed an additional lower stage of production.

[6]By well-behaved f we mean a continuous, twice differentiable function which, in addition, yields a monotonically decreasing function with respect to q_{wi} on the left-hand side of (10).

cases discussed below:

Case 1: *Cournot Oligopoly.* Consider l horizontally independent downstream firms which integrate vertically with l horizontally independent upstream firms where $l \leq \min(m,n)$. The first-order profit-maximization conditions for these integrated firms is then

$$(5') \quad \alpha(f + f'\alpha q_{wi}) = c_i \qquad i = 1,2,\ldots,l$$

where the c_i's are the marginal costs of the upstream firms.[7] These costs are independent of, and conceived to be lower than the p_w of equation (5). Except for this specification, equation (5′) is equivalent to (5).

Now, the marginal revenues provided on the left-hand side of (5) and (5′) are the same for each of the downstream firms regardless of whether or not they themselves integrate with an upstream supplier. Moreover, the profit-maximizing condition for the $m - l$ nonintegrated firms remains as given before by (5) for these $i = l + 1, \ldots, m$ firms. (These $m - l$ dealers can be conceived for simplicity to purchase their gasoline tankwagons from the $n - l$ independent (oligopolistic) refineries.) Aggregating (5′) and (5), respectively, over $i = 1,2,\ldots,l$ and $i = l+1,\ldots,m$ therefore establishes

$$(6') \qquad \alpha\left(f + \frac{1}{l}f'\alpha q_w^l\right) = \frac{1}{l}\sum c_i, \qquad q_w^l = \sum_{i=1}^{l} q_{wi}$$

and

$$(6'') \quad \alpha\left(f + \frac{1}{m-l}f'\alpha q_w^m\right) = p_w, \qquad q_w^m = \sum_{i=l+1}^{m} q_{wi}$$

These two equations, respectively, provide the aggregated inputs of the l integrated and $m - l$ nonintegrated downstream producers. Note in particular that equation (6″) represents the market demand for the gasoline tankwagons in the perspective of the $n - l$ nonintegrated upstream firms. Profit maximization by these input suppliers involves

$$(9a) \quad \alpha\left(g^\dagger + \frac{1}{m-l}f''\alpha^2 q_w^m q_{wj} + \frac{m-l+1}{m-l}f'\alpha q_{wj}\right) = c_j \qquad j = l+1,\ldots,n$$

where $g^\dagger = f + [1/(m-l)]f'\alpha q_w^m$. Since profit maximization by the l vertically integrated refineries is already given by (5′), it follows that in order to derive the counterpart to (10), we simply have to sum (9a) over $j = l+1,\ldots,n$ to obtain

$$(10a) \quad \alpha\left(g^\dagger + \frac{1}{(m-l)(n-l)} \cdot f''\alpha^2 q_w^m q_w^n + \frac{m-l+1}{(m-l)(n-l)}f'\alpha q_w^n\right) = \frac{1}{n-l}\sum c_j \qquad q_w^n = \sum_{j=l+1}^{n} q_{wj}$$

Market equilibrium requires the additional equations

$$(11) \qquad q_w^n = q_w^m$$

$$(12) \qquad q_w = q_w^n + q_w^l$$

These equations, together with (6′), (6″), and (10a), can be solved simultaneously for the five endogenous variables, q_w, q_w^n, q_w^m, q_w^l, p_w, where the derived equilibrium input price p_w must exceed the marginal costs (c_j) of the independent upstream firms.

A significant conclusion derives from the requirement $p_w > c_j$. Combining (6′) and (6″) towards this end, we obtain

$$(13) \quad \alpha g = p_w - (1/m)\sum_{j=1}^{l}(p_w - c_j) < p_w$$

where the correspondent to f, to recall, is defined as $g = f + (1/m)f'\alpha q_w$. If, in contrast, no vertical integration has taken place at all, equation (6) applies, repeated below for convenience as

[7]Vertical integration per se may reduce the costs of production, c's. In addition, economies of scale may operate to reduce the c's since integration increases output. We will assume away all of these possibilities, and concentrate our attention on the economics of integration that are related directly to market structures

(13′) $\alpha g = p_w$

Identity of the left-hand side of (13) and (13′), as well as the condition (pursuant to fn. 6) that they are decreasing functions of q_w, are each well established above. However, the middle term of (13) is less than the right-hand side of (13′).[8] The equilibrium result q_w is, therefore, unambiguously greater under the conditions of equation (13) than that which underscores (13′). The supply (q_L as well as q_w) must, in other words, increase with vertical integration, and the final price p_L must decline in accordance with equation (2).[9]

Case 2: Stackelberg Oligopoly. Intrinsic to the foregoing argument was the basic assumption that the l firm(s) which integrate vertically with the l oligopolistic input supplier(s) would behave *with respect to the noncolluded firms as would the firms in Cournot's world of oligopoly.* This assumption is, however, not requisite to our basic welfare results. In fact, if it were not for possible antitrust prosecution, the firms which integrate vertically with input suppliers could be expected not only to identify the reaction functions of noncolluded firms, but to *lead them* accordingly. The same general final market solution given previously would nevertheless obtain, as demonstrated below.

Consider Stackelberg's leader-follower market, and let l final good producers collude vertically with input suppliers. These firms, we propose, take the nonintegrated dealer's reaction functions as data in fulfilling the requirements of Stackelberg's leader-follower oligopoly. Their profit-maximization conditions would therefore not be given by (5′) but by

$$(5'') \quad \frac{d\pi_i}{dq_{wi}} = 0, \quad \alpha\Big(f + f'\alpha q_{wi} \cdot \Big[1 + \frac{dq_w^m}{dq_{wi}}\Big]\Big) = c_i \qquad i = 1, 2, \ldots, l$$

where dq_w^m/dq_{wi} stems from the reaction functions of the $(m - l)$ nonintegrated firms. These reaction functions, provided by (5), are repeated below as

$$(14) \quad \alpha(f + f'\alpha q_{wj}) = p_w, \qquad j = l + 1, \ldots, m$$

Differentiating (14) with respect to q_{wi}, while holding *all other merged firms' supply* as well as p_w constant, generates

$$(15) \quad \alpha\Big(f'\Big[1 + \frac{dq_w^m}{dq_{wi}}\Big] + f'\frac{dq_{wj}}{dq_{wi}} + f''\alpha q_{wj}\Big[1 + \frac{dq_w^m}{dq_{wi}}\Big]\Big) = 0, \qquad j = l + 1, \ldots, m$$

Aggregating for all j's and rearranging terms establishes

$$(16) \quad \frac{dq_w^m}{dq_{wi}} = \frac{f'}{f'(m - l + 1) + f''\alpha q_w^m} - 1$$

Substituting (16) back into (15) yields

$$(17) \quad \alpha\Big(f + \frac{f'^2\alpha q_{wi}}{f'(m - l + 1) + f''\alpha q_w^m}\Big) = c_i \qquad i = 1, 2, \ldots, l$$

which equation provides the profit-maximizing conditions for the l merged firms.

[8] This condition requires stable levels of P_w before and after vertical integration when l is small relative to m. However, such stability is not even needed as l approaches m, i.e., as more and more firms integrate vertically.

[9] Perry argued that vertical integration requires profit incentives among the merging firms. However, such merger does not require greater industry profits. Rather, additional mergers would be feasible if they increase the profits of the merging firms at the expense of previously merged or still remaining independent firms. To appreciate the above proposition, assume $\alpha = 1$, $f(q) = a - q$, and let $m = n = 2$. Then it can readily be shown that vertical integration increases the profit of the initial merged firms, i.e., where $l = 1$. Incentive exists accordingly for the integration. However, the profit for the remaining independent firms, and the industry profit as well, can be shown to decrease in the process. There exist incentives, in other words, for the excluded (or slowpoke) firms to integrate vertically to recover their lost profits. In fact, the final result with $l = 2$ is characterized by lower total industry profit than that which prevailed initially when no integration at all had occurred, i.e., when $l = 0$. Merged firms would now find the original state a better one. Can they go back to *their* garden of Eden? Not individually without simultaneous agreement.

This equation contrasts with (5′), the profit-maximizing equation for the $(m - l)$ independent firms. Combining (17) and (5′) for all firms produces the counterpart to (13), namely

$$(13'') \quad \alpha\left(f + \frac{1}{m} f'\left[\alpha q_w^m + \frac{f'}{[f'(m - l + 1) + f''\alpha q_w^m]}\alpha q_w^l\right]\right) = p_w - \frac{1}{m}\sum (p_w - c_i)$$

where the major parenthesis term of (13″) remains above the correspondent g whenever the major bracketed term is less than αq_w. The sufficient condition for this particular relation is $f'' \leq 0$, although even if $f'' > 0$, the left-hand side of the above formula would continue to lie above g so long as $f'/[f'(m - l + 1) + f''\alpha q_w^m]$ is always less than unity. Under these general (rather nonrestrictive) conditions on the shape of the market demand function f, comparison of (13″) with (13) indicates sharply that market equilibrium supply would be even greater under vertical collusion in a Stackelberg world than under vertical collusion in a Cournot world of oligopoly.

III. Conclusion

The basic thesis of this paper applies in general to oligopolistic industries in which fixed proportions apply to successive stages of production. We propose, accordingly, that for industries so characterized, integration should be allowed without concern over arbitrary concentration ratios. In reverse pattern, our purely micro-economic inquiry into the subject of size indicates, for example, that the breaking up of large oil companies in the United States, *ceteris paribus*, would tend to diminish output.

REFERENCES

A. Cournot, *Mathematical Principles of the Theory of Wealth*, trans. N. T. Bacon, New York 1927.

M. L. Greenhut and H. Ohta, "Related Market Conditions and Interindustrial Mergers," *Amer. Econ. Rev.*, June 1976, *66*, 257–77.

——— **and** ———, "Related Market Conditions and Interindustrial Mergers: Reply," *Amer. Econ. Rev.*, Mar. 1978, *68*, 228–30.

J. R. Haring and D. L. Kaserman, "Related Market Conditions and Interindustrial Mergers: Comment," *Amer. Econ. Rev.*, Mar. 1978, *68*, 225–27.

G. A. Hay, "An Economic Analysis of Vertical Integration," *Ind. Org. Rev.*, No. 3, 1973, *1*, 188–98.

F. Machlup and M. Taber, "Bilateral Monopoly, Successful Monopoly, and Vertical Integration," *Economica*, May 1960, *27*, 101–19.

M. K. Perry, "Vertical Integration of Successive Monopolists: Comment," *Amer. Econ. Rev.*, Mar. 1978, *68*, 221–24.

R. Schmalensee, "A Note on the Theory of Vertical Integration," *J. Polit. Econ.*, July/Aug. 1972, *80*, 442–49.

J. M. Vernon and D. A. Graham, "Profitability of Monopolization by Vertical Integration," *J. Polit. Econ.*, July/Aug. 1971, *79*, 924–25.

F. R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, July/Aug. 1974, *82*, 783–801.

Charitable Contributions: New Evidence on Household Behavior

By WILLIAM S. REECE*

In this paper I provide some new empirical evidence on the impact of tax deductibility on the level of personal charitable contributions. In order to provide empirical evidence of this kind we have to consider the broader question of the determination of the level of charitable giving by the household. The "price of contributions," which is dependent upon the tax treatment of those contributions, is only one element of this process. Efforts to quantify the impact of tax deductibility lead to an empirical model which allows us to test a number of hypotheses on the determination of the level of charitable contributions.

In Section I, a model of personal charitable giving and its empirical implications are discussed. In Section II previous attempts to test some of these implications and to quantify the impact of tax deductibility on charitable giving are discussed. Section III presents the specification of the empirical model including a discussion of the data, and Section IV presents empirical results. In Section V the implications of these results are summarized.

I. A Model of Personal Charitable Giving

Economic theory is generally concerned with exchange, or two-way transfer of economic goods. Philanthropic behavior, or the voluntary one-way transfer of economic goods to individuals or organizations outside the family unit, has been rationalized in the economics literature by the hypothesis that individuals' preferences are defined over levels of consumption of unrelated persons as well as levels of their own consumption.[1] Kenneth Boulding and William Vickrey were among the first modern economists to suggest this rationalization of charitable giving. Gary Becker later provided a formal model of this kind of behavior and used it to derive some empirical implications of the utility interdependence hypothesis.

Very briefly, Becker's model is based on the maximization, subject to the individual's budget constraint, of a utility function representing the individual's preferences over the levels of his own consumption and the consumption of others about whom the individual is concerned. The level of consumption of others is assumed to have two components: the "social environment," which is the level of consumption of others in the absence of contributions by the individual, and the individual's contribution. Becker's model implies the individual's optimal level of contributions varies directly with his income, and inversely with the price of contributions and the level of consumption of others in the absence of contributions. Further, his model, along with the assumption that the consumption of others is a luxury from the donor's viewpoint, implies that the income elasticity of demand for contributions ex-

*Research coordinator, Management Information Division, University of Minnesota. The research presented here was conducted while I was an economist in the Division of Price and Index Number Research of the U.S. Bureau of Labor Statistics, and as part of my duties there. An earlier version of this paper was distributed as *BLS* Working Paper No. 76. Without implicating them in any of the shortcomings of this paper, I would like to thank Theodore Bergstrom, Steven Cobb, Arthur Denzau, Robert Gillingham, Dennis Sullivan, William Weiler, the managing editor of this *Review*, and an anonymous referee for comments on an earlier draft. Special thanks are due to Herbert Cover whose programming assistance made this project possible.

[1]Philanthropic behavior has also been rationalized as the result of the existence of an "alliance" among the members of the society. (See Louis De Alessi.) In this view, maintaining the organization of society is a collective good toward which individuals will contribute.

ceeds unity.[2] The price of contributions is less than unity because many charitable contributions are deductible when computing personal income tax liability. Thus, less than one dollar of own consumption is foregone with each one dollar of contributions. Prices of own-consumption goods also influence the optimal level of contributions, but the directions of influence are made ambiguous by conflicting income and substitution effects.

II. Earlier Empirical Models of Household Charitable Behavior

A. *Summary of Results*

There have been a number of attempts in recent years to provide some empirical evidence on the determinants of the level of household charitable contributions. Michael Taussig, using a cross section of individual tax return data, found income elasticities greater than unity and a very small and statistically insignificant price elasticity. Robert Schwartz used a time-series of aggregate tax return data and found both income and price elasticities to be less than unity in absolute value. Schwartz attempted to test the interdependence hypothesis by including in his equation per capita nondonor income as a proxy for the consumption of the relevant recipient group. This variable had a highly significant negative coefficient. Harold Hochman and James Rodgers (1973), using average income and contributions data for thirty-two metropolitan areas from the 1960–61 *Survey of Consumer Expenditures* conducted by the Bureau of Labor Statistics (*BLS*), estimated the income elasticity of demand for contributions to be greater than unity. They did not include a price variable in the equation. They also tested a variant of the interdependence hypothesis by including in the equation a variable measuring the dispersion of income within the metropolitan area. This variable had a highly significant coefficient with the predicted positive sign. Martin Feldstein (1975a), using a time-series of aggregate tax return data, estimated the income elasticity to be less than unity and the price elasticity to be greater than unity. In "Part II" (1975b) he used a cross section of data on charitable contributions by income class to different types of institutions. He found income elasticities exceeding unity for contributions to education institutions and hospitals and less than unity for contributions to health and welfare institutions and to religious organizations. The price elasticities for all categories except religious contributions were found to be greater than unity in absolute value. Feldstein and Amy Taylor used cross sections of tax return data (including information on state income tax rates) and estimated the income elasticity to be somewhat less than unity and the price elasticity to be greater than unity. Feldstein and Charles Clotfelter, using a cross section of household data from the 1963–64 Board of Governors survey of financial characteristics, found income elasticities of less than unity and price elasticities exceeding unity. Michael Boskin and Feldstein used data from the 1974 *National Study of Philanthropy* conducted by the Survey Research Center of the University of Michigan and found price elasticities substantially exceeding unity and income elasticities somewhat less than unity.

In summary, the results for equations using all contributions as the dependent variable tend to point to an income elasticity of demand for contributions of less than unity and a price elasticity exceeding unity.[3] Further, the limited evidence avail-

[2]Thus, an income elasticity of demand for contributions in excess of unity is not an implication of the interdependent utility hypothesis alone. Rather, it is an implication of the hypothesis in combination with this assumption about the income elasticity of demand for the consumption of others.

[3]The major exception to this is Taussig. Feldstein and Taylor rationalize the contrasting results they obtain using the same basic data source as Taussig. First they point out that Taussig inadvertently used only two-thirds of the complete sample, resulting in a potentially biased subsample. Second, they argue that his income and price variables were inappropriate. Taussig used the marginal tax rate for taxable income net of contributions; that is, the price of the last dollar of contributions. Feldstein and Taylor

able supports the utility interdependence hypothesis. Finally, the results showing different price and income elasticities for different types of contributions (Feldstein 1975b) indicate that it may be inappropriate to model the determination of aggregate contributions with a single equation.

B. *Problems with Earlier Empirical Models*

The results discussed in the previous section have provided valuable evidence on household charitable behavior. However, the empirical specifications previously used are each subject to question. Perhaps the most serious problem with many of the previous studies is their reliance on data obtained from tax returns. This restricts the sample to those who itemize their deductions, eliminating most of the households on the low end of the income spectrum. With the exception of Feldstein (1975b), all the previous studies have used total contributions as their dependent variable rather than breaking down contributions by type of recipient. Thus, we might suspect they have aggregated various types of contributions which should be modeled separately. (For example, Vickrey has argued that many religious contributions can be thought of as payments for support of religious services used by the donor rather than as truly philanthropic payments.) Further, we might expect the level of charitable contributions to be determined by permanent income rather than current measured income, yet only Schwartz and Feldstein and Clotfelter made any attempt to use permanent income. Finally, real income effects and substitution effects caused by interregional cost-of-living variation are ignored in all the cross-section studies.

III. The Empirical Model

The new data to be used, a subset of the 1972–73 *BLS Consumer Expenditure Survey* (*CEX*), allows us to avoid to some extent the problems discussed above. These data consist of extensive surveys of the incomes, expenditures and personal characteristics for a large sample of households. The data on previous year's income contained in the *CEX* will allow us to attempt to measure permanent income. We can also identify the state and Standard Metropolitan Statistical Area (*SMSA*) of residence (as well as location within the *SMSA*) of the households which will enable us to allow for state and local tax rate differences and make some allowance for interregional cost-of-living variation. Also the *CEX* contains data on eight classifications of contributions, so that we can see whether or not the same empirical model is applicable to contributions to different types of recipients.

A. *Contributions*

The *CEX* contains data on eight classifications of contributions:

- *SUPPORT:* cash contributions for support of persons not in the consumer unit (including alimony)
- *GIFTS:* gifts of cash, bonds or stocks to persons not in the consumer unit
- *CHARITY:* contributions to charities, such as the United Fund, Red Cross, etc., which were not deducted from pay
- *RELIGIOUS:* contributions to church and other religious organizations, excluding parochial school expenses
- *EDUCATIONAL:* contributions to educational organizations

point out that this variable is dependent upon the level of contributions and thus introduces a spurious positive correlation between price and the level of giving, biasing the price coefficient toward zero. In order to eliminate this problem they calculate the price of contributions using the marginal tax rate for the level of taxable income assuming no charitable deduction. (I calculate the price variable in the same way.) Taussig's income variable was income net of taxes paid. Feldstein and Taylor argue the correct variable should be income net of the taxes that would have been paid if there had been no charitable deduction, once again because this variable will not depend upon the level of contributions. However, there also seems to be a problem with this income variable since it is possible for contributions plus expenditures on goods to exceed income defined in this way. That is, using this definition of income the budget constraint is endogenous. The correct definition of income to be used as the independent variable in the contributions equation thus seems to be gross income.

POLITICAL: political contributions
DEDUCTED: contributions to charities deducted from pay
OTHER: other

Some of these classes of contributions are of questionable philanthropic nature. Contributions for support of persons not in the household are questionable because of the inclusion of alimony under that category. Gifts of cash or securities to persons not in the household will include intergenerational transfers within the family. Contributions to religious organizations may be payments for services. Political contributions are related to the welfare of others only in a very indirect way, and may be a form of investment. Also, although contributions to educational institutions are usually thought of as philanthropic, it is not clear that the motivations of the contributors are related to a concern for the levels of consumption of others.[4] If we exclude all these questionable classes of contributions our primary contributions variable would be *CHARITY* + *DEDUCTED*. However, we will also see how well the model performs using all contributions as well as each of the eight classes individually.

B. *Prices*

I define the price of a dollar of contributions to be the amount of own consumption foregone by making the contribution. Frequently this will be strictly less than unity because of the income tax deductibility of many charitable contributions. For tax deductible contributions by those who itemize their deductions the price of an additional dollar of cash contributions is $(1 - T_m)$, where T_m is the marginal income tax rate.[5] For nondeductible contributions and all contributions made by those taking the standard deduction, the price of contributions is unity. Since we have no direct information on which households in our sample itemized their tax deductions we infer which households itemized by estimating each household's deductions and comparing this estimate with the standard deduction.[6] We assume those households with the standard deduction exceeding the itemized deductions did not itemize and assign them a price of contributions equal to unity. All others are assigned prices equal to $1 - [(1 - T_{ms})T_{mf} + T_{ms}]$, where T_{mf} is the marginal federal tax rate and T_{ms} is the marginal state and local tax rate for the households' taxable incomes.[7]

In order to account for differences in the prices of own consumption goods I include in our equation a variable denoted *COL*, representing differences in prices faced by the households. It is constructed from the *BLS* intermediate family budget indexes for 1972 and 1973.

C. *Own Income*

I experimented with both current and permanent income concepts in the contributions equations. Following Feldstein and Clotfelter, I used the average of current

[4]Further, to the extent that they are purchases of a memorial to the donor, they are not philanthropic. However, in our sample containing no extremely wealthy individuals, this kind of contribution is not likely to be a factor.

[5]Contributions of appreciated assets will have a lower price due to their favorable tax treatment. Data limitations force us to ignore this complication. See Feldstein (1975a) for a detailed discussion of the price of contributions of appreciated assets.

[6]The standard deduction is calculated as 15 percent of taxable income with a $1,300 minimum and a $2,000 maximum. Itemized deductions are estimated to be the sum of all interest payments, state and local taxes paid and medical deductions. Medical deductions are set equal to one-half of medical insurance payments (up to $150) plus the amount of medical expenses (including the remainder of medical insurance payments) exceeding 3 percent of federal taxable income. For the purposes of discriminating between itemizers and nonitemizers, deductible charitable contributions are also added to deductions. (When calculating the marginal tax rates these are deleted.) Households with itemized deductions exceeding the standard deduction are assumed to have itemized.

[7]We have a model in which income determines (in part) the price, and income and price both affect the level of contributions. Income and price are not rigidly tied together, however, since some households have nontaxable income (especially the return to homeownership), and there is variation in marginal tax rates not due to income because of the variation in state and local tax rates. The correlation between *PRICE* and *INCOME* is −.63.

year and previous year incomes to represent permanent income. Since the results were not sensitive to the choice of income variable and permanent income consistently outperformed current income, I report only the permanent income results. In order to reduce the importance of transitory components in the income variable, the sample is restricted to nonfarm households in which the household heads are male, are the only earners, are not self-employed, and were employed at least 26 weeks during the survey year.

I was also able to add to permanent income an estimate of the net returns to homeownership for those households which own their own homes. Homeowning households were asked to estimate the monthly rental values of their homes. From these estimates, I subtracted monthly mortgage interest, ground rent, and home taxes and added this annualized net amount to permanent income.[8]

D. *Social Environment*

This variable presents the most difficult data problem for this study because the relevant potential recipient groups are not known for the households in our sample. To make this variable operational, assume each household is most concerned about the consumption levels of those households geographically close to it.[9] I was able to obtain unpublished census data on the distributions of income within the twenty largest *SMSA*s for the year 1975. Restricting the sample to these twenty *SMSA*s, I use an approximation of lower quintile income for the *SMSA* within which the household resides to represent the level of consumption of potential recipients.[10] Since public assistance payments can substitute for recipient income we also include average public assistance payments for the *SMSA* within which the donor household resides as an additional variable representing the social environment.[11]

E. *Other Variables*

Feldstein and Clotfelter suggest that the age of the household head may be an important additional determinant of the level of charitable contributions. I include age of the household head as an additional independent variable to test this hypothesis. I also include a dummy variable *SECOND* which takes on the value unity for those observations from the 1973 sample and zero for those from the 1972 sample. To summarize, the independent variables in our contributions equation, in addition to a constant, are

PRICE: price of contributions (defined above)

INCOME: average of current and previous years' family income before taxes plus net return from homeownership

ASSISTANCE: average public assistance

[8]Homeowners were also asked to estimate the current market values of their homes. For those who would not estimate the rental value but did estimate the market value, I used 1 percent of the latter in place of the rental value. Those who did not give either estimate are dropped from the sample. If the calculated net return to homeownership is negative I set the value to zero.

[9]This is an important assumption upon which the validity of the results is based. More realistically we might assume the donor's "synthetic family" is much more narrowly defined than this; for example, it may be restricted by religion as well as location. Unfortunately data limitations preclude finer delineations of the potential recipient group.

[10]In addition to the restrictions on the sample enumerated above, I eliminated those households with real estate sales during the survey year because of the lack of data on the extent to which these sales represent capital gains. Also eliminated from the sample was one observation judged to be an outlier. This household had extremely large contributions to the *CHARITY* classification, amounting to approximately one-third of its gross income. This is in contrast to the average *CHARITY/INCOME* ratio of .0025. This ratio for the outlier household is twenty-two standard deviations from the mean. Further, when this observation is included in the sample the mean squared residual for equation (1), Table 1, rises to 296,293 from 8402, and the predicted value for the outlier observation is only 5 percent of the actual value. The final sample has 537 households.

[11]Hochman and Rodgers (1969) consider the theory of the use of the public sector to achieve income redistribution goals which result from utility interdependence.

(Aid to Families with Dependent Children, *AFDC*) + old age assistance + aid for the physically and totally disabled per recipient in the *SMSA* in which the household resides (Feb. 1973, annualized)

RECIPIENT: lower quintile family income for the *SMSA* within which the household resides (1975)

COL: intermediate family budget index for the *SMSA* in which the household resides (for 1972 observations the 1972 family budget index is multiplied by 1.057 to adjust for the change in the Consumer Price Index (*CPI*) from July 1972 to July 1973)

AGE: age of the household head

SECOND: unity for those observations from the 1973 sample and zero for those observations from the 1972 sample.

IV. Empirical Results

I estimate the elasticities of various types of contributions with respect to the independent variables listed in Section III by means of a maximum likelihood Tobit technique.[12] Tables 1 and 2 show the elasticities implied by the estimated coefficients of the Tobit index function, the coefficients themselves, and their t-values. The tables also show two measures of goodness of fit of each of the equations. The traditional R^2 is inappropriate because the mean errors of the equations are nonzero. The two measures shown are the alternatives suggested by Henri Theil, $1 - e'e/s^2$ and $1 - e'e/y'y$, where $e'e$ is the sum of squared residuals, s^2 is the variance of the dependent variable, and $y'y$ is the second moment about zero of the dependent variable.[13]

Table 1 shows these results for the primary contributions variable *CHARITY + DEDUCTED*, the sum of all eight classes *ALL*, and the sum of all except *SUPPORT*, *GIFTS*, and *OTHER* called *CONTRIB*, which approximates the dependent variable used in previous studies of this type using tax return data (with the exception of Feldstein, 1975b). The equation with *CHARITY +*

TABLE 1—ESTIMATED CONTRIBUTIONS EQUATIONS FOR THE PRIMARY DEPENDENT VARIABLES

	Equations (1) *CHARITY + DEDUCTED*	(2) *ALL*	(3) *CONTRIB*
PRICE	−.976	−1.401	−1.192
	−114.60	−787.88	−396.71
	(−2.67)	(−4.63)	(−4.15)
INCOME	1.423	.550	.877
	.0095	.0176	.0166
	(9.99)	(4.87)	(8.01)
AGE	.309	.484	.380
	.8808	6.60	3.06
	(1.44)	(2.79)	(2.30)
ASSISTANCE	−.097	−.186	.102
	−.0108	−.0996	.0322
	(−.29)	(−.67)	(.39)
RECIPIENT	−.138	.327	.351
	−.0017	.0190	.0121
	(−.37)	(1.06)	(1.20)
COL	−1.511	.518	−.542
	−.1420	.2329	−.1443
	(−1.21)	(.51)	(−.57)
SECOND	−.016	.005	−.012
	−3.42	5.32	−7.30
	(−.27)	(.11)	(−.26)
CONSTANT	124.70	113.33	183.61
	(.95)	(.22)	(.64)
$1 - e'e/s^2$	.342	.175	.282
$1 - e'e/y'y$	.466	.405	.529

Note: The elasticity, coefficient, and t-statistic (in parentheses) are given for each variable.
Sources: U.S. Bureau of Labor Statistics except *RECIPIENT*, U.S. Bureau of the Census, and *ASSISTANCE*, U.S. Department of Health, Education, and Welfare.

[12] The Tobit model is based on the assumption that for each household there exists an index I which is a linear function of the variables explaining the level of the dependent variable and a random error term:

$$I_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_n X_{ni} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

When $I_i \leq 0$ the value of the dependent variable is set to zero. When $I_i > 0$ the dependent variable is set equal to I_i. (See James Tobin for a more detailed and general presentation of the Tobit model.) The implied elasticity of the expected value of the dependent variable, $E(Y)$, with respect to an independent variable, X_i, is $\beta_i X_i F(X\beta/\sigma)/E(Y)$, where $F(\ )$ is the cumulative normal distribution function. (See Robert Shishko and Bernard Rostker, p. 303, for the derivation of the elasticity.)

[13] See Theil, pp. 175–78, especially fn. 4.

TABLE 2—ESTIMATED CONTRIBUTIONS EQUATIONS FOR THE REMAINING DEPENDENT VARIABLES

	Equations: (4) SUPPORT	(5) GIFTS	(6) CHARITY	(7) RELI-GIOUS	(8) EDUCA-TIONAL	(9) POLITI-CAL	(10) DEDUCTED	(11) OTHER
PRICE	-5.279	-.139	-.402	-1 598	- 077	-2 962	-2 707	-3 066
	-5787 34	-21 71	-33.41	-457 72	-7 75	-116 74	-286 61	-130.60
	(-2 92)	(-.25)	(- 99)	(-4 44)	(- 06)	(-1 80)	(-3.53)	(-1 03)
INCOME	- 738	420	1 668	397	1 639	959	355	.002
	- 0461	0037	0079	0065	.0094	.0022	.0021	0000
	(-1.19)	(2 08)	(10 48)	(2 96)	(3 89)	(1 96)	(1 32)	(00)
AGE	1 076	986	166	467	.678	.927	638	2 109
	28.59	3 73	3337	3 24	1 66	8861	1 64	2 18
	(1 23)	(3 17)	(70)	(2 29)	(.82)	(1 00)	(1 46)	(1 49)
ASSISTANCE	-2 201	124	026	102	-1 196	.754	- 288	1 185
	-2 30	0184	0020	0278	- 1153	0284	- 0291	0481
	(-1.51)	(23)	(07)	(31)	(- 87)	(51)	(- 44)	(48)
RECIPIENT	314	1 004	- 119	566	1 919	109	- 039	- 108
	0355	.0162	- 0010	0167	0200	0004	- 0004	- 0005
	(19)	(1 77)	(-.28)	(1 54)	(1 34)	(07)	(- 05)	(- 04)
COL	4 380	6 324	- 014	158	8 016	-3 321	-6.407	-15 384
	3 84	7985	- 0010	0362	6482	- 1047	-.5428	- 5243
	(83)	(3 40)	(- 01)	(13)	(1 62)	(- 61)	(-2 49)	(1 45)
SECOND	- 038	-.204	- 089	- 012	079	- 493	131	435
	-75 59	-57.80	-13 46	-6 29	14 61	-35 38	25 29	33 69
	(- 15)	(-2 30)	(-1 34)	(- 21)	(35)	(-1 84)	(1 07)	(96)
CONSTANT	-1535 12	-1221 94	-83 87	55 54	-1287 34	-12 78	613 48	304 87
	(- 30)	(-4 59)	(- 81)	(18)	(-2 89)	(- 07)	(2 51)	(76)
$1 - e'e/s^2$	036	047	375	116	- 037	- 030	114	008
$1 - e'e/y'y$	056	172	475	362	- 017	- 004	189	016

Note See Table 1
Sources See Table 1.

DEDUCTED as the dependent variable fits better than that using *ALL* by both goodness-of-fit measures and better than that using *CONTRIB* by one measure.[14] These results are in accord with the a priori expectation that the utility interdependence model would not model the contributions of questionable philanthropic nature as well as it models *CHARITY* + *DEDUCTED*.

The *PRICE* and *INCOME* variables perform quite well in equation (1), Table 1. Their coefficients are of the expected signs and are quite significantly different from zero. The price elasticity is very close to unity and the income elasticity exceeds unity. Those variables intended to represent the consumption of potential recipients, *ASSISTANCE* and *RECIPIENT*, have insignificant, negative coefficients. Thus, the utility interdependence hypothesis does not receive much support from the results. However, since *ASSISTANCE* and *RECIPIENT* only roughly approximate the theoretical concepts they are intended to represent, these results should not be regarded as conclusive. The coefficients of the remaining variables in equation (1), *AGE*, *SECOND* and *COL*, are insignificant.

In the equations with *ALL* and *CONTRIB* as dependent variables one sees somewhat larger price elasticities and smaller income elasticities. These values are remarkably close to those estimated in the earlier studies using a comparable dependent variable. It can also be seen that in the equations using the aggregate dependent variables the coefficient on *AGE* becomes statistically significant.

In Table 2 the results of using each of the eight individual classes of contributions as the dependent variable are shown. The goodness-of-fit measures, estimated elasticities, and *t*-values vary widely depending

[14]The *CONTRIB* equations fit better than the *CHARITY* + *DEDUCTED* equations by the second goodness-of-fit measure because *CONTRIB* has a much larger mean than *CHARITY* + *DEDUCTED* ($201.87 vs. $54.53), and $y'y$ exceeds s^2 by the square of the mean of the dependent variable.

upon the dependent variable. The best fit is obtained when *CHARITY* is used as the dependent variable. The goodness-of-fit measures are extremely low for the equations for *SUPPORT*, *EDUCATIONAL*, *POLITICAL*, and *OTHER*. Indeed, in the equations for *EDUCATIONAL* and *POLITICAL* the second moments about zero of the residuals exceed the second moments of the dependent variables. Goodness of fit as measured by $1 - e'e/s^2$ is also very low for the *GIFTS* equation. Thus, these types of contributions are definitely not well modeled by my empirical specification.

In Table 2 the coefficient of *AGE* is statistically significant in only two equations, those with *GIFTS* and *RELIGIOUS* as the dependent variables.[15] The coefficient of *INCOME* is generally significant, but its elasticity is essentially zero in the *OTHER* equation and is negative in the *SUPPORT* equation. These results help explain the differences in elasticities between equations (1), (2), and (3). The increase in size and significance of the *AGE* elasticity as the level of aggregation increases seems to be due to the inclusion of *RELIGIOUS* in *CONTRIB* and the inclusion of both *RELIGIOUS* and *GIFTS* in *ALL*. Also the inclusion of *SUPPORT* and *OTHER* in *ALL* seems to be responsible for the low income elasticity and high price elasticity for the *ALL* equation, while the inclusion of *RELIGIOUS* in *CONTRIB* seems to have reduced the income elasticity in the equation explaining the latter. Thus, the aggregation of different kinds of contributions, which my results suggest cannot be modeled by a single empirical specification, has important implications for the estimated elasticities.[16] This suggests that one should be cautious in using parameters estimated using aggregate contributions variables in simulating the effects of alternative tax treatments on aggregate contributions.

It can be seen in Table 2 that *ASSISTANCE* and *RECIPIENT* do not perform well. Further, the coefficient of *COL* is quite erratic across equations and is generally insignificant.[17] Using a one-tailed t-test at the .95 level, the coefficient of *SECOND* is significant and negative in the *POLITICAL* equation, which is reassuring since 1972 was an election year. This variable is also negative and significant in the *GIFTS* equation.

Some of the *PRICE* coefficients are rather surprising, especially when compared to the earlier results obtained by Feldstein (1975b). He found a price elasticity of only .49 for contributions to religious organizations while mine is almost 1.6. He also found a highly significant price elasticity of 2.23 for contributions to educational institutions while my results fail to show such a relationship. These contrasting results should be reconciled before one attempts to simulate the distributional effects of tax law changes.[18]

V. Conclusions

1) The utility interdependence hypothesis receives little support from my results.

[15]These results support the hypotheses that intergenerational transfers and religious contributions increase as death approaches

[16]The reader may feel there is an inconsistency in my arguing against aggregation while using *CHARITY + DEDUCTED* as the primary variable in spite of the differences in the equations for the two variables taken separately. However, the distinction between *CHARITY* and *DEDUCTED* is artificial from the point of view of the model I am modeling contributions by *type of recipient*. The distinction between *CHARITY* and *DEDUCTED* is not by type of recipient, but rather by *means of payment*, a subject about which the model has nothing to say. The equations for *CHARITY* and *DEDUCTED* taken separately are included for completeness, not because they are meaningful for my purposes.

[17]The *COL* results are consistent with the earlier assertion that, because of conflicting income and substitution effects, the sign of the coefficient of the price of own-consumption goods might be either positive or negative However, the results might more reasonably be attributed to the fact that the *COL* variable only roughly approximates the concept it is intended to represent and to the potential specification error in the equations for those contributions other than *CHARITY + DEDUCTED*.

[18]These contradictory results may perhaps be explained by the differences in samples. Feldstein's data are more heavily weighted with upper income households which make the majority of the education contributions but relatively little of the religious contributions My sample, on the other hand, consists almost entirely of households with annual incomes under $40,000.

The estimated income elasticity for the primary dependent variable exceeds unity, as my model, augmented by the assumption that the consumption of others is a luxury, suggests. The coefficients of the variables used to represent the levels of consumption of potential recipients have the correct sign but are insignificant. This latter result cannot be considered strong evidence against the utility interdependence hypothesis, however, since the levels of consumption of the potential recipients of each household's contributions are only roughly approximated.

2) My results indicate that the tax deductibility of charitable contributions is an important determinant of the level of the primary contributions variable, which represents contributions to health and welfare organizations (such as the United Fund and the Red Cross). The estimated unitary price elasticity indicates that these organizations gain one dollar of contributions for each dollar of tax revenue lost because of the deductibility of these contributions.[19]

3) My results indicate that religious organizations gain even more than is lost in tax revenue because of tax deductibility. This last result should be treated with caution, however, since it conflicts with earlier evidence obtained by Feldstein (1975b) using quite different data. Additional research should be conducted with data sets covering a broader spectrum of households in order to identify the source of the conflict.

4) My model has a very poor fit for most of the dependent variables (representing contributions to different kinds of recipients). The major exception is the primary dependent variable. Further, my estimated price and income elasticities are sensitive to the level of aggregation of the dependent variable. These results indicate that a single empirical specification should not be used to explain the level of contributions aggregated by type of recipient.

[19]See Feldstein and Taylor, p. 1207.

REFERENCES

G. S. Becker, "A Theory of Social Interactions," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1063–93.

M. J. Boskin and M. S. Feldstein, "Effects of the Charitable Deduction on Contributions by Low Income and Middle Income Households: Evidence from the National Survey on Philanthropy," *Rev. Econ. Statist.*, Aug. 1977, *59*, 351–54.

K. E. Boulding, "Notes on a Theory of Philanthropy," in Frank G. Dickinson, ed., *Philanthropy and Public Policy*, New York 1962, 57–71.

L. De Alessi, "Toward a Theory of Postdisaster Cooperation," *Amer. Econ. Rev.*, Mar. 1975, *65*, 127–38.

M. Feldstein, (1975a) "The Income Tax and Charitable Contributions: Part I—Aggregate and Distributional Effects," *Nat Tax J.*, Mar. 1975, *28*, 81–100.

———, (1975b) "The Income Tax and Charitable Contributions: Part II—The Impact on Religious, Educational and Other Organizations," *Nat. Tax J.*, June 1975, *28*, 209–26.

——— **and C. Clotfelter,** "Tax Incentives and Charitable Contributions in the United States: A Microeconomic Analysis," *J Publ. Econ.*, Jan./Feb. 1976, *5*, 1–26.

——— **and A. Taylor,** "The Income Tax and Charitable Contributions," *Econometrica*, Nov. 1976, *44*, 1201–22.

H. M. Hochman and J. D. Rodgers, "Utility Interdependence and Income Transfers Through Charity," in Kenneth E. Boulding et al., eds., *Transfers in an Urbanized Economy*, Belmont 1973, 63–77.

——— **and** ———, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–57.

R. A. Schwartz, "Personal Philanthropic Contributions," *J. Polit. Econ.*, Nov./Dec. 1970, *78*, 1264–91.

R. Shishko and B. Rostker, "The Economics of Multiple Job Holding," *Amer. Econ. Rev.*, June 1976, *66*, 298–308.

M. K. Taussig, "Economic Aspects of the Personal Income Tax Treatment of Charitable Contributions," *Nat. Tax J.*, Mar. 1967, *20*, 1–19.

Henri Theil, *Principles of Econometrics*, New York 1971.

J. Tobin, "Estimation of Relationships for Limited Dependent Variables," *Econo-*

metrica, Jan. 1958, *26*, 24–36.

W. S. Vickrey, "One Economist's View of Philanthropy," in Frank G. Dickinson, ed., *Philanthropy and Public Policy*, New York 1962, 31–56.

U.S. Bureau of the Census, unpublished data.

U.S. Bureau of Labor Statistics (*BLS*), *Consumer Expenditure Survey*, unpublished data (on tape).

U.S. Department of Health, Education, and Welfare, *Recipients of Public Assistance Money Payments and Amounts of Such Payments, By Program, State, and County, February 1973*, Washington 1973.

Optimal Financing of the Government's Budget: Taxes, Bonds, or Money?

By ELHANAN HELPMAN AND EFRAIM SADKA*

In this paper we investigate the optimal financing of the government's budget. In particular, we consider taxation, bond issuance, and money creation as alternative means of financing.

These issues are not new. Some of them were extensively explored in the public finance literature; some of them were explored in the monetary theory literature. What we believe to be new, however, is that our work integrates the relevant elements from both strands of literature into a unified general equilibrium framework. With this approach we are able to jointly determine the optimal combination of tax rates, the rate of inflation, and the interest rate on government bonds.

This type of analysis enables us to ask whether the rate of inflation should be positive or negative, and to provide simple conditions on compensated demand elasticities which determine the answer. It also enables us to reconsider Milton Friedman's optimum quantity of money rule in a public finance framework without lump sum taxes, and to show how this rule should be modified. In addition, we provide insight into the desirability of a social security scheme by showing that such a program is undesirable when social security payments are financed from general revenue.[1]

We try to keep things as simple as possible in order to concentrate on the main issues. In Section I we develop a model of overlapping generations in which money serves only as a store of value and it is the only store of value. The optimal steady-state financing of the government's budget for this model is analyzed in Section II. We then show in Section III that the results of Section II (properly interpreted) remain valid when money provides additional services. Government bonds are also introduced in Section III. Friedman's optimum quantity of money rule is then reconsidered. Throughout the analysis we assume that there is no population growth. Implications of a positive rate of population growth are explored in our working paper.

I. A Simple Model

Consider a consumption-loan model of identical individuals (see Paul Samuelson) in which every individual lives for two periods. In the first period of his life an individual works, consumes, and saves; all savings are done in the form of money holdings (in Section III the model is extended to include bonds). In the second period he retires and consumes the fruits of his first-period savings. The rate of population growth is assumed to be zero (see, however, our working paper for a discussion of a positive rate of population growth). For simplicity we also assume that each generation consists of one individual.

We assume that there is only one consumption good and that it is produced by means of labor with a fixed coefficient technology. The labor-output ratio is assumed to be one. This implies that the gross wage rate is equal to the producer price of the consumption good whenever the good is produced.

The government is assumed to purchase in each period a fixed quantity of the consumption good. Its expenditure is financed

*Department of economics, Tel-Aviv University and University of Rochester, and department of economics, Tel-Aviv University and University of Wisconsin, respectively. An early version of this paper was presented at the 1977 Hanuka Meeting of the Israel Economic Association. We wish to thank the participants of the Faculty Seminar at Tel-Aviv University who hotly debated with us some issues raised in this paper. In particular we would like to thank Jacob Frenkel and Yoram Weiss.

[1]The work of Edmund Phelps is the closest in spirit to the present study. He did, however, concentrate on other issues.

by means of (a) commodity taxes; (b) labor income taxes; and (c) deficit finance; that is, printing money.[2] *The government is not allowed to impose lump sum taxes.*

In the most general formulation that follows, we assume that the consumption of young people can be taxed at a different rate than the consumption of old people. Restrictions on the government's ability to use discriminatory taxation schemes is discussed in the sequel.

A. *Individuals*

Consider a young individual at time x. He has a standard utility function $u(c_1, c_2, L)$, where $c_1(\geq 0)$ is his first-period consumption, $c_2(\geq 0)$ is his second-period consumption, and L (≤ 0) is his first-period labor supply. (In Section III we consider the case in which $u(\cdot)$ depends also on real balance holdings.) Assuming that he correctly predicts prices and wages in period $x + 1$, his consumption and labor supply programs are restricted by the following first- and second-period budget constraints:

$$(1)\quad (1 + \tau_1)p(x)c_1 + (1 - \theta)w(x)L + M(x + 1) = 0$$

$$(2)\quad (1 + \tau_2)p(x + 1)c_2 - M(x + 1) = 0$$

where

τ_i = consumption tax rate for an individual of age i, $i = 1, 2$

θ = tax rate on labor income

$p(x)$ = producer price of the consumption good in period x

$w(x)$ = gross wage rate in period x (the wage rate paid out by the producer)

$M(x)$ = stock of money held *at the beginning* of period x, $M(x) \geq 0$

Since the labor-output ratio equals one, $p(x) = w(x)$ for every x (assuming that production takes place).

In a steady state per capita real balances are constant, $M(x)/p(x) = m$ for every x, and the rate of inflation $\pi = [p(x + 1)/p(x)] - 1$ is constant and equal to the rate of growth of the nominal per capita stock of money. Hence, in a steady state (1) and (2) can be written as

$$(3)\quad (1 + \tau_1)c_1 + (1 - \theta)L + (1 + \pi)m = 0$$

$$(4)\quad (1 + \tau_2)c_2 - m = 0$$

In a steady state a young individual maximizes his utility function subject to (3), (4), an upper limit on labor supply, and $m \geq 0$. If the individual chooses $m = 0$, deficit finance is impossible.

In a general model positive savings cannot be assured unless one makes assumptions about the time preference and about the relative scarcity of the second-period endowment as, for example, in Jean-Michel Grandmont and Guy Laroque. In our case, it is sufficient to assume that second-period consumption is indispensable (i.e., its marginal utility goes to infinity as c_2 goes to zero), since the individual cannot work in the second period of his life. In any case, we assume that for the range of tax rates and inflation rates considered by us, there is voluntary positive savings; that is, the solution to the consumer problem implies positive money balance holdings. In this case it is legitimate to combine (3) and (4) into a single budget constraint by eliminating m, resulting in the following consumer problem:

(5) Choose $c_1 \geq 0, c_2 \geq 0, \bar{L} \leq L \leq 0$, to maximize $u(c_1, c_2, L)$

subject to

$$(6)\quad (1 + \tau_1)c_1 + (1 + \pi)(1 + \tau_2)c_2 + (1 - \theta)L = 0$$

where $-\bar{L}$ is the first-period bound on labor supply.

This is a standard consumer problem in which $(1 + \tau_1)$ is the consumer price of c_1,

[2]We can distinguish between deficit finance and inflationary finance. By deficit finance we mean a situation in which the stock of money is changed by means of deficits or surpluses in the government's budget. By inflationary finance we mean a situation in which the financing of the government's budget causes inflation or deflation. The two concepts differ only when the rate of population growth differs from zero.

$(1+\pi)(1+\tau_2)$ is the consumer price of c_2, and $(1-\theta)$ is the consumer price of L.[3]

The meaning of these prices is quite simple. Consider a young consumer who lives in period x. For present-period consumption he faces the price $p(x)(1+\tau_1)$, the producer price plus the indirect tax on young consumers. For labor services he faces the price $w(x)(1-\theta) = p(x)(1-\theta)$, the gross wage rate minus the wage tax. In order to buy a consumption good in the second period of his life he has to save $p(x+1)(1+\tau_2)$, the producer price plus the indirect tax on old consumers. Hence, from the young individual's point of view, the price of second-period consumption is $p(x+1)(1+\tau_2) = p(x)(1+\pi)(1+\tau_2)$. Now, since only relative prices matter, divide all prices by $p(x)$ to obtain the price vector $[1+\tau_1, (1+\pi)(1+\tau_2), (1-\theta)]$.

The solution to (5)-(6) yields ordinary demand functions $c_1(\cdot)$, $c_2(\cdot)$, and $L(\cdot)$, and an indirect utility function $v(\cdot)$, all having consumer prices as their arguments. Thus, the following is a complete representation of the indirect utility function:

$$(7)\quad v = v[1+\tau_1, (1+\pi)(1+\tau_2), 1-\theta; 0]$$

where the last argument stands for nonwage income which equals zero.

B. *The Government*

In every period x, the government buys $G > 0$ units of the consumption good. Its budget constraint is

$$(8)\quad p(x)G = \tau_1 p(x)c_1 + \tau_2 p(x)c_2 - \theta w(x)L + M(x+1) - M(x)$$

The first three components on the right-hand side of (8) represent tax proceeds, while the last two represent money injection, the deficit finance component. In the steady state (using $p(x) = w(x)$), (8) can be written as

$$(9)\quad G = \tau_1 c_1 + \tau_2 c_2 - \theta L + \pi m$$

The last term on the right-hand side of (9) represents revenue from inflation. It equals the rate of inflation multiplied by the real stock of money, evaluated at producer prices. Thus, π can be considered as a tax on real balances.

Substituting (4) into (9), we obtain

$$(10)\quad G = \tau_1 c_1 + [\tau_2 + \pi(1+\tau_2)]c_2 - \theta L$$

Substituting the demand functions into (10), the optimal choice of tax rates and the rate of inflation can be represented by

(11) Choose $\tau_1, \tau_2, \theta, \pi, t_1, t_2, t_3$

to maximize $v(1+t_1, 1+t_2, 1+t_3; 0)$

subject to

$$(12)\quad t_1c_1(1+t_1, 1+t_2, 1+t_3; 0) + t_2c_2(1+t_1, 1+t_2, 1+t_3; 0) + t_3L(1+t_1, 1+t_2, 1+t_3; 0) - G = 0$$

and

$$(13a)\quad t_1 = \tau_1$$

$$(13b)\quad t_2 = \tau_2 + \pi(1+\tau_2)$$

$$(13c)\quad t_3 = -\theta$$

Observe that (11) and (13) are consistent with (7) and that (12) and (13) are consistent with (10). In order to make sure that problem (11)-(13) takes properly into account the production constraint, combine (6) with (12) using (13) to obtain

$$(14)\quad G + c_1(\cdot) + c_2(\cdot) + L(\cdot) = 0$$

which assures production feasibility. In fact

[3]It is also clear from the budget constraint (6) that in a steady state the real rate of return on savings is

$$-\frac{dc_2}{dc_1} - 1 = \frac{1+\tau_1}{(1+\pi)(1+\tau_2)} - 1 = \frac{\tau_1-\tau_2}{(1+\tau_2)(1+\pi)} - \frac{\pi}{1+\pi}$$

Hence, commodity taxes as well as the rate of inflation determine the real rate of return of savings. If, however, commodity taxes do not discriminate between young and old consumers (i.e., $\tau_1 = \tau_2$), the real rate of return on savings depends only on the rate of inflation and is equal to $-\pi/(1+\pi)$, which—for small rates of inflation—is approximately equal to $-\pi$.

(14) can replace (12) in the optimization problem.[4]

II. The Optimal Tax and Inflation Rates

Problem (11)–(13) can be solved in two stages. In the first stage one can choose $t = (t_1, t_2, t_3)$ to maximize $v(\cdot)$ subject to (12); in the second stage one can choose $(\tau_1, \tau_2, \theta, \pi)$ to satisfy (13) for the optimal values of the t_i's, $t^* = (t_1^*, t_2^*, t_3^*)$. Moreover, if t^* is a solution, then—since the indirect utility function and the demand functions are homogeneous of degree zero in prices and income—for every $\lambda > 0$, $[\lambda(1 + t_1^*) - 1, \lambda(1 + t_2^*) - 1, \lambda(1 + t_3^*) - 1]$ is also an optimal tax scheme. Hence, we may assume that one t_i is zero. The second stage has a solution because for every vector $t = (t_1, t_2, t_3)$ there exist $(\tau_1, \tau_2, \theta, \pi)$ which satisfy (13). Moreover, for every t there exists $(\tau_1, \tau_2, \theta, \pi)$ with $\pi = 0$ which satisfies (13). This proves:

PROPOSITION 1: *If the consumption of young people can be taxed at a different rate than the consumption of old people, then inflationary finance is a redundant policy instrument. Inflationary finance can be replaced by commodity taxation with no change in welfare.*[5]

The intuition behind this proposition is quite clear. All that inflation or deflation does is to change the ratio between the first-period price and the second-period price of consumption (see problem (5)–(6)). Hence, if other taxes can perform this discrimination, there is no need to use inflation or deflation.

Suppose now that it is impossible to impose differential commodity tax rates on young and old consumers—either due to arbitrage considerations or for political reasons. This means that we have the additional constraint $\tau_1 = \tau_2 = \tau$. Hence, unless $t_1^* = t_2^*$, inflationary finance is desirable. For if $t_1^* - t_2^* \neq 0$, (13a) and (13b) imply $\tau_1 - \tau_2 - \pi(1 + \tau_2) = -\pi(1 + \tau) \neq 0$, which can be satisfied only if $\pi \neq 0$. In this case $(\tau_1^*, \tau_2^*, \theta^*, \pi^*) = \{t_1^*, t_1^*, -t_3^*, (t_2^* - t_1^*)/(1 + t_1^*)\}$ is the solution. Clearly, in this case inflation is desirable if $t_2^* > t_1^*$; i.e., if it is optimal to tax old peoples' consumption (or second-period consumption) at a higher rate than young peoples' consumption (or first-period consumption); deflation is desirable if it is optimal to tax young peoples' consumption at a higher rate than old peoples' consumption. Inflationary finance should not be used if optimality requires uniform taxation of the consumption good.

In a general optimal taxation model it is hard to provide conditions which enable the comparison of optimal tax rates on different goods. However, in a three-commodity world such comparison is possible. In particular, Wilfred J. Corlett and Douglas C. Hague and Peter Diamond and James Mirrlees showed that in a three-commodity model the good which has the higher compensated demand elasticity with respect to the price of the numeraire should be taxed at a lower rate. If both taxed goods have the same compensated demand elasticity with respect to the price of the numeraire, optimality requires a uniform tax rate on both taxed goods. Using labor as a numeraire, this proves the following proposition:

PROPOSITION 2: *Suppose we have the additional constraint $\tau_1 = \tau_2$. Then inflation is desirable if the compensated demand elasticity of first-period consumption with respect to the wage rate is higher than the compensated demand elasticity of second-period consumption with respect to the wage rate, and deflation is desirable in the opposite case. If the two compensated demand elasticities are equal, inflationary finance should not be used.*

[4]In this way we can interpret our optimization problem as that of choosing consumer prices $(1 + t_1, 1 + t_2, 1 + t_3)$ so as to maximize $v(\cdot)$ subject to (14). In this formulation the producer prices and wage rate are all one.

[5]Recall that we have assumed zero population growth, which means that inflationary finance is equivalent to deficit finance. See the authors for the case of nonzero population growth.

It seems reasonable that first-period consumption is a better substitute for first-period leisure than second-period consumption. In this case a positive rate of inflation is desirable. If, however, the utility function is weakly separable in first- and second-period consumption, that is, if $u(c_1, c_2, L) \equiv z[f(c_1, c_2), L]$ and $f(\cdot)$ is homothetic, then inflationary finance should not be used. The reason is that in this case the relevant compensated demand elasticities are equal (see, for example, Agnar Sandmo). This proves:

PROPOSITION 3: *Suppose we have the constraint $\tau_1 = \tau_2$. Then if $u(c_1, c_2, L) \equiv z[f(c_1, c_2), L]$ and $f(\cdot)$ is homothetic, then inflationary finance should not be used.*

So far we have identified cases in which inflationary finance is not required, should not be used, or cases in which inflation or deflation is the optimal policy. This raises the following question: in which cases should we not finance the *entire* government budget by means of a monetary expansion?

PROPOSITION 4: *If c_1 and c_2 are net substitutes in the Hicks-Slutzky sense, then it is not optimal to use only deficit financing.*

PROOF:

Suppose to the contrary, that $\tau_1^* = \tau_2^* = 0$ and $\pi^* \neq 0$ is an optimal solution. Then $\pi^* > 0$ since in the present case $G = \pi^* m$ and $G, m > 0$. This means that $t^* = (0, \pi^*, 0)$ is a solution to the first-stage problem in (11). Let c_{ij}^o denote the partial derivative of the compensated demand for consumption in period i with respect to price j, where $i = 1, 2$ and $j = 1, 2, 3$. The optimal tax formulae can be written as (see Diamond and Mirrlees, p. 263, equation (42)):

(a) $$t_1^* c_{11}^o + t_2^* c_{12}^o = -\alpha^* c_1$$

(b) $$t_1^* c_{21}^o + t_2^* c_{22}^o = -\alpha^* c_2$$

where $\alpha^* \geq 0$. Since $t_1^* = 0$ and $t_2^* = \pi^* > 0$, (i) implies $c_{12}^o \leq 0$, which means that c_1 and c_2 are ***not*** net substitutes.

We expect first- and second-period consumption to be net substitutes. Therefore, we expect some degree of taxation to be optimal. It can also be shown that it is not optimal to use only deficit financing when either c_1 or c_2 are net complements with L. But this relationship seems to us less interesting than substitutability between c_1 and c_2.

Finally, let us consider a social security scheme as a complementary policy tool. Suppose the government pays out to every old person $S(x)$ dollars in period x. Suppose also that these payments are fully indexed so that in a steady state $S(x + 1) = (1 + \pi) \cdot S(x)$. In this case social security payments consist of fixed real per capita transfers from the government to old consumers. These are lump sum transfers; but we do not allow lump sum taxes.

Now we can ask: what is the optimal composition of taxes, inflation, and social security payments? The answer is that in the present context *optimal security payments are zero;* that is, it is optimal to have no social security system. The reason is that in this case social security payments—which are lump sum transfers—have to be financed by distortionary taxation. It is therefore desirable to reduce the lump sum transfers as much as possible, namely to zero. (For a formal proof of the proposition that it is always desirable to reduce lump sum transfers which are financed by distortionary taxes see the Appendix in Anthony Atkinson and Nicholas Stern.)

III. On Money and Bonds

So far we assumed that money serves only as a store of value. However, monetary theory attributes to money additional functions, like the provision of liquidity services and saving of transaction costs. This means that money provides utility to its holders above the value of the store of value services. It is customary to summarize the utility from these functions by specifying a utility index which includes real money balances as one of its arguments. Suppose, therefore, that our utility function is

(15) $$U = U(c_1, c_2, L, m_c)$$

where m_c are real balances in terms of *consumer* prices that an individual holds at the beginning of the second period of his life. There are other possible specifications, but for our illustrative purposes this one seems to be as good as any other.

There is a simple way in which the previous results can be reinterpreted to apply to the present framework. In a steady state $m_c = m/(1 + \tau_2)$, where $m = M(x)/p(x)$, and the individual chooses $m/(1 + \tau_2) = c_2$ from (4). Hence, defining

$$u(c_1, c_2, L) \equiv U(c_1, c_2, L, c_2) \tag{16}$$

we see that our discussion in Sections I and II is valid in the present framework if $u(\cdot)$ is interpreted as the reduced form of $U(\cdot)$, as in (16).

There is, however, one important issue with which we cannot deal by employing the above respecification, and this is the issue of government borrowing and lending. When money serves only as a store of value, there is no point in issuing government bonds, since they play in this setting the same role as money. But if money also provides other services, there is room in the system for both money and bonds. Then we can ask: what is the optimal interest rate to be paid on government bonds? For simplicity, we let the government choose the interest rate, allowing the public to both borrow from and lend to the government at this interest rate.

The answer to this question sheds light on Friedman's optimum quantity of money rule. The rule says that, since the marginal cost of producing real balances is zero, it is optimal to increase real balances to the point at which their marginal utility is driven to zero; that is, to the point at which there is satiation in real balance holdings. This can be done by driving to zero the nominal interest rate, which will make money a perfect store of value substitute for bonds. Alternatively, it is proposed to pay interest on money (see Harry Johnson) in order to obtain the same effect.

In our framework Friedman's rule is satisfied if it is *optimal* to pay zero interest on government bonds. We show, however, that this is an unlikely result. The reason is that our framework of analysis, which is in the spirit of what has become to be known as the New Public Economics, considers financing of the government's budget as a *second best* problem, in which case there is no room for lump sum taxation

Let $B(x)$ be the stock of the government's one-period bonds that the consumer wants to hold at the beginning of period x ($B(x)$ may be negative). $B(x)$ is measured in nominal terms; that is, in dollars. Let r be the nominal interest rate. Assuming at the outset that young consumers have to be taxed at the same rate as old consumers (i.e., $\tau_1 = \tau_2 = \tau$), constraints (1) and (2) become

$$(1 + \tau)p(x)c_1 + (1 - \theta)w(x)L + B(x + 1) + M(x + 1) = 0 \tag{17}$$

$$(1 + \tau)p(x + 1)c_2 - (1 + r) \cdot B(x + 1) - M(x + 1) = 0 \tag{18}$$

In a steady state per capita real balances $m = M(x)/p(x)$, as well as per capita real government debt $b = B(x)/p(x)$, are constant; the stock of money and the government's nominal debt grow at the rate of inflation. Hence, in the steady state (17) and (18) can be written as

$$(1 + \tau)c_1 + (1 - \theta)L + (1 + \pi)b + (1 + \pi)(1 + \tau)m_c = 0 \tag{19}$$

$$(1 + \tau)c_2 - (1 + r)b - (1 + \tau)m_c = 0 \tag{20}$$

Since there is no sign restriction on b (the government may either borrow from or lend to the public), these two budget constraints can be combined into one. Multiplying (20) by $(1 + \pi)/(1 + r)$ and adding the result to (19), we obtain

$$(1 + \tau)c_1 + \frac{(1 + \tau)(1 + \pi)}{1 + r} c_2 + (1 - \theta)L + \frac{r(1 + \tau)(1 + \pi)}{1 + r} m_c = 0 \tag{21}$$

Constraint (21) is a young consumer's budget constraint. He maximizes $U(c_1, c_2, L, m_c)$ subject to (21), which yields the indirect utility function

$$(22)\quad V = V[1+\tau, (1+\tau)(1+\pi)/(1+r), 1-\theta, r(1+\tau)(1+\pi)/(1+r); 0]$$

The consumer price of first-period consumption is $1+\tau$, the consumer price of second-period consumition is $(1+\tau)\cdot(1+\pi)/(1+r)$, the consumer price of labor supply and leisure is $1-\theta$, and the consumer price of real balances m_c is $r(1+\tau)\cdot(1+\pi)/(1+r)$. Only the prices of second-period consumption and of real balances m_c require an interpretation; the price of first-period consumption and the price of leisure (labor supply) have not changed, and their interpretation is the same as in Section II.

In order to buy a unit of the consumption good in the second period of his life, a young consumer who lives in period x needs $p(x+1)(1+\tau) = p(x)(1+\pi)(1+\tau)$ dollars in the second period of his life. Since he can save by purchasing government bonds, which yield an interest rate r, he has to save in the first period of his life $p(x)(1+\pi)(1+\tau)/(1+r)$ dollars. This explains the price of c_2 in (21), since we normalize by dividing all prices by $p(x)$.

Now, in order to get one unit of real balances m_c at the beginning of his second period of life, the young consumer in period x has to save $p(x)(1+\tau)(1+\pi)$ dollars and to *keep* it in the form of money. This causes him an interest loss of $p(x)r(1+\tau)(1+\pi)$ dollars which could have materialized had he saved in bonds. The present value of this interest loss is $p(x)r(1+\tau)(1+\pi)/(1+r)$. Hence, his price of real balance holdings, after the normalization in which we divide all prices by $p(x)$, is $r(1+\tau)(1+\pi)/(1+r)$.

It is also easy to see that in this case the real rate of return on money holdings is $-\pi/(1+\pi)$ and the real rate of return on bond holdings is $(r-\pi)/(1+\pi)$. Hence, if the rate of inflation equals the nominal interest rate, the real rate of return on bond holdings is zero.

The government's budget constraint (see (8)) now becomes

$$(23)\quad rB(x) + p(x)G = \tau p(x)(c_1+c_2) - \theta p(x)L + B(x+1) - B(x) + M(x+1) - M(x)$$

The left-hand side represents interest payments on the government's outstanding debt plus spending on the purchase of goods. The first three components on the right-hand side represent tax revenue, and they are followed by terms which represent the increase in the government's debt and the monetary injection. In a steady state this becomes

$$(24)\quad G = \tau(c_1+c_2) - \theta L + (\pi - r)b + (1+\tau)\pi m_c$$

Using (20), (24) becomes

$$(25)\quad G = \tau c_1 + \left[\frac{(1+\tau)(1+\pi)}{1+r} - 1\right]c_2 - \theta L + \frac{r(1+\tau)(1+\pi)}{1+r}m_c$$

Equation (25) describes the government's budget constraint in terms of real tax rates. This can be seen by comparing (25) with the consumer prices which enter the indirect utility function (22). Recall from (22) that the consumer is "buying" first- and second-period consumption, real balances m_c, and he is selling labor services. These are his relevant commodities; bonds have no direct value. The normalized consumer prices of c_1, c_2, L and m_c, are $(1+\tau)$, $(1+\tau)(1+\pi)/(1+r)$, $(1-\theta)$, and $r(1+\tau)(1+\pi)/(1+r)$, respectively.

Now, the normalized producer prices of c_1, c_2, L, and m_c, are 1, 1, 1, and 0, respectively. Consumption and labor have the same producer price because a unit of labor was assumed to produce a unit of output and from the production point of view consumption of old and young consumers are perfect substitutes. Real balances are "produced" with zero marginal (and average) costs. Hence, the producer price of real balances is zero. Using this interpretation the reader can see that the right-hand side of (25) represents real taxes collected by the government, where the difference between

consumer and producer prices serve as the tax rates. These tax rates are the real tax rates implicit in the economy; they are compounds of the government's tax and financial policies.

Thus, maximization of (22) subject to (25), where demand functions replace c_1, c_2, L, and m_c in (25), yields a standard optimal taxation problem. It can be solved, as before, in two stages. In the first stage we choose t_1, t_2, t_3, and t_4 to maximize

$$V(1 + t_1, 1 + t_2, 1 + t_3, t_4; 0) \tag{26}$$

subject to

$$\begin{aligned} & t_1c_1(1 + t_1, 1 + t_2, 1 + t_3, t_4; 0) \\ &+ t_2c_2(1 + t_1, 1 + t_2, 1 + t_3, t_4; 0) \\ &+ t_3L(1 + t_1, 1 + t_2, 1 + t_3, t_4; 0) \\ &+ t_4m_c(1 + t_1, 1 + t_2, 1 + t_3, t_4; 0) = G \end{aligned} \tag{27}$$

and in the second stage we choose τ, θ, r, and π to satisfy

$$\tau = t_1 \tag{28a}$$

$$\theta = -t_3 \tag{28b}$$

$$r = t_4/(1 + t_2) \tag{28c}$$

$$\pi = (t_2 + t_4 - t_1)/(1 + t_1) \tag{28d}$$

The question that we address now is whether the optimum quantity of money rule proposed by Friedman is valid when there is no way to raise revenue by means of nondistorting instruments. If it turns out to be the case that the optimal interest rate on government bonds is zero (or $t_4 = 0$), then Friedman's optimum quantity of money rule is also optimal in the present framework. We show, however, that under reasonable conditions the optimal interest rate is *positive*, which means that at the optimum individuals are not satiated in real balance holdings, contrary to Friedman's rule. The reason for this result is that in a second best framework it may be optimal to tax a commodity with zero marginal costs of production, and a positive interest rate implies the existence of a positive tax rate on real balance holdings.

Denoting the optimal solution to the tax problem by an asterisk, we prove:

PROPOSITION 5: *Let c_1 and c_2 be net substitutes, and let each one of them be a net substitute for real balance holdings m_c, all in the Hicks-Slutzky sense. Then the optimal interest rate on bonds is positive.*

PROOF:

Let, without loss of generality, L be the numeraire (i.e., we set $t_3 = 0$). Then the optimal tax formulae can be written as (see Diamond and Mirrlees, p. 263, eq. (42))

$$\text{(a)} \quad t_1^*c_{11}^o + t_2^*c_{12}^o + t_4^*c_{14}^o = -\alpha^*c_1$$

$$\text{(b)} \quad t_1^*c_{21}^o + t_2^*c_{22}^o + t_4^*c_{24}^o = -\alpha^*c_2$$

$$\text{(c)} \quad t_1^*m_{c1}^o + t_2^*m_{c2}^o + t_4^*m_{c4}^o = -\alpha^*m_c$$

where $\alpha^* \geq 0$ and c_{ij}^o, m_{cj}^o are compensated demand derivatives.

Suppose $t_4^* = 0$. Then (a), (b), and (c) imply

$$\text{(a}'\text{)} \quad t_1^*c_{11}^o + t_2^*c_{12}^o \leq 0$$

$$\text{(b}'\text{)} \quad t_1^*c_{21}^o + t_2^*c_{22}^o \leq 0$$

$$\text{(c}'\text{)} \quad t_1^*m_{c1}^o + t_2^*m_{c2}^o \leq 0$$

Since $G > 0$, we have either $t_1^* > 0$ and/or $t_2^* > 0$, in order to raise revenue. However, (a′) and (b′) imply $t_1^*, t_2^* > 0$. For suppose $t_1^* \leq 0$, in which case $t_2^* > 0$. This contradicts (a′) since $c_{11}^o \leq 0$ (nonpositive own-price effect), and $c_{12}^o > 0$ by assumption. Similarly suppose $t_2^* \leq 0$ in which case $t_1^* > 0$. This contradicts (b′) since $c_{22}^o \leq 0$ and $c_{21}^o > 0$. Hence, $t_1^*, t_2^* > 0$. This, however, contradicts (c′), since $m_{c1}^o, m_{c2}^o > 0$, by assumption. Hence, $t_4^* \neq 0$. Since t_4^* is the consumer price of m_c, it must be the case that $t_4^* \geq 0$ and hence $t_4^* > 0$. Since $1 + t_2^*$ is the consumer price of c_2 in which there is no satiation (by assumption), then $1 + t_2^* > 0$. Hence, it follows from (28c) that $r^* > 0$.

Thus we have seen that in a wide class of problems the optimal interest rate is positive. In these cases Friedman's optimum quantity of money rule is *not* optimal.

In the case in which money serves only as a store of value, we saw that a nonzero rate of inflation is desirable only if it is desirable to make the consumer price of second-period consumption in terms of

first-period consumption different from one. In the present context this is no longer the case. In particular, if it is optimal to have the consumer price of second-period consumption in terms of first-period consumption equal to one, then under the conditions of Proposition 5 it is optimal to have a positive rate of inflation. For this price ratio equals one if and only if $t_1^* = t_2^*$, which together with (28c,d) implies $\pi^* = r^*$, and r^* was proved to be positive.

Finally, observe that the presence of bonds does not change our argument concerning the desirability of a social security system which is financed from general revenue. Making lump sum payments to old individuals is undesirable, because the revenue needed for these payments has to be raised by means of distorting instruments.

REFERENCES

A. B. Atkinson and N. H. Stern, "Pigou, Taxation and Public Goods," *Rev. Econ. Stud.*, Jan. 1974, *41*, 119–28.

W. J. Corlett and D. C. Hague, "Complementarity and the Excess Burden of Taxation," *Rev. Econ. Stud.*, No. 1, 1953, *21*, 21–30.

P. A. Diamond and J. A. Mirrlees, "Optimal Taxation and Public Production: II," *Amer. Econ. Rev.*, June 1971, *61*, 261–78.

M. Friedman, "The Optimum Quantity of Money," in his *The Optimum Quantity of Money and Other Essays*, Chicago 1969.

J. M. Grandmont and G. Laroque, "Money in the Pure Consumption Loan Model," *J. Econ. Theory*, Aug. 1973, *6*, 382–95.

E. Helpman and E. Sadka, "Optimal Financing of the Government's Budget: Taxes, Bonds, or Money?," work. paper no. 18-77, Foerder Instit. Econ. Res., Tel-Aviv Univ., June 1977.

H. G. Johnson, "Is There an Optimal Money Supply?," *J. Finance*, May 1970, *25*, 435–42.

E. S. Phelps, "Inflation in the Theory of Public Finance," *Swedish J. Econ.*, Mar. 1973, *75*, 67–82.

P. A. Samuelson, "An Exact Consumption Loan Model of Interest With or Without the Social Contrivance of Money," *J Polit. Econ.*, Dec. 1958, *66*, 467–82.

A. Sandmo, "A Note on the Structure of Optimal Taxation," *Amer. Econ. Rev.*, Sept. 1974, *64*, 701–06.

Alternative Theories of Pricing, Distribution, Saving, and Investment

By HANS BREMS*

The whole dispute between Keynesian and non-Keynesian theories is whether investment determines savings, or vice versa. *Nicholas Kaldor* [1966]

The purpose of this paper is to identify some similarities and dissimilarities between neoclassical and post-Keynesian models of growth and distribution. Four questions will be asked. First, does saving or investment adjust to a higher propensity to save? Second, within that adjustment, what are the roles played by the real wage rate and mark-up pricing? Third, within that adjustment, does a Wicksell Effect emerge? Fourth, are the models examined open or closed, and if closed, how?

Answers will be facilitated by specifying both models mathematically, in their stripped form, and in the same notation:

Variables:

C ≡ consumption
g_v ≡ proportionate rate of growth of variable v ≡ S and X
I ≡ investment
κ ≡ physical marginal productivity of capital stock
L ≡ labor employed
m ≡ mark-up pricing factor
P ≡ price of good
S ≡ physical capital stock
W ≡ wage bill
X ≡ physical output
Y ≡ national money income
Z ≡ profits bill

Parameters:

a, b ≡ input-output coefficients
α, β ≡ exponents of Cobb-Douglas production function
c ≡ propensity to consume
F ≡ available labor force
g_F ≡ proportionate rate of growth of parameter F
M ≡ multiplicative factor of Cobb-Douglas production function
w ≡ money wage rate

I. Equations Common to Both Models

We confine ourselves to the stripped form of either model having one good, an immortal capital stock of that good, and no technological progress in it. Four definitions and one equilibrium condition are common to both models. Define the proportionate rate of growth:

(1) $$g_v \equiv \frac{dv}{dt}\frac{1}{v}$$

Define investment as the time derivative of capital stock:

(2) $$I \equiv \frac{dS}{dt}$$

Define the wage bill as the money wage rate *times* employment:

(3) $$W \equiv wL$$

Define national money income as the sum of wage and profits bills:

(4) $$Y \equiv W + Z$$

Equilibrium requires output to equal demand for it:

(5) $$X = C + I$$

II. Equations Peculiar to the Neoclassical Model

A. *Production*

Let entrepreneurs apply a Cobb-Douglas production function

(6) $$X = ML^{\alpha}S^{\beta}$$

*Professor of economics, University of Illinois at Urbana-Champaign. For critical comments, I am grateful to George H. Borts, an anonymous referee of this *Review*, Carl G. Uhr, and Poul Nörregaard Rasmussen.

where $0 < \alpha < 1$, $0 < \beta < 1$, and $\alpha + \beta = 1$. Profit maximization under pure competition will equalize the real wage rate and the physical marginal productivity of labor:

$$\frac{w}{P} = \frac{\partial X}{\partial L} = \alpha \frac{X}{L} \tag{7}$$

Define physical marginal productivity of capital as

$$\kappa \equiv \frac{\partial X}{\partial S} = \beta \frac{X}{S} \tag{8}$$

Multiply κ by the value of capital stock PS and define profits as

$$Z \equiv \kappa PS = \beta PX \tag{9}$$

Assume full employment:

$$F = L \tag{10}$$

B. *Distributive Shares*

Insert (7) into (3), (3) and (9) into (4), and find $Y = PX$ and the distributive shares $W/Y = \alpha$ and $Z/Y = \beta$. Notice that this result is derived before specifying the neoclassical consumption function

$$C = cX \tag{11}$$

where $0 < c < 1$. In the distributive shares, then, the form of the consumption function makes no difference. For example, the latter may well be of the post-Keynesian variety like (18) below with realistic different propensities to consume real wages and real profits. In that case we would simply have $c \equiv \alpha c_W + \beta c_Z$.

C. *Does Saving or Investment Adjust to a Higher Propensity to Save?*

If the propensity to save $1 - c$ were twice as high, how would a neoclassical model adjust? The adjustment would begin with a typical response of factor proportion to relative factor price. Divide (6) by L and (7) by α and set the resulting expressions for X/L equal. Find

$$S/L = (\alpha M)^{-1/\beta}(w/P)^{1/\beta} \tag{12}$$

So a higher real wage rate w/P would induce a higher capital intensity S/L. But what would the real wage rate be?

D. *The Real Wage Rate and the Wicksell Effect*

To solve for the real wage rate, divide (6) by S, raise both sides to the power -1, and find the capital coefficient:

$$S/X = (1/M)(S/L)^{\alpha} \tag{13}$$

Next use (1) and (2) to write $I \equiv g_S S$, insert that and (11) into (5), and find another expression for the capital coefficient:

$$S/X = (1 - c)/g_S$$

Finally set the right-hand sides of the two expressions for S/X equal, insert the result into (12), and find the real wage rate:

$$w/P = \alpha M^{1/\alpha}[(1 - c)/g_S]^{\beta/\alpha} \tag{14}$$

Here is the Wicksell Effect: According to (14) the real wage rate w/P is higher the higher is the propensity to save $1 - c$ or, as Knut Wicksell himself expressed his effect: "The capitalist saver is, thus, fundamentally, the friend of labour" (p. 164).

E. *Closing the Neoclassical Model*

Equation (14) is not a solution yet. So far it merely expresses one unknown, the real wage rate w/P, in terms of another, the rate of growth of capital stock g_S. But neoclassicists do close their models. In the absence of technological progress they find proportionate rates of growth to be converging to the steady-state solutions:

$$g_S = g_X = g_F \tag{15}$$

Insert (15) into (14), and you have a neoclassical solution for the real wage rate w/P.

The full-adjustment mechanism is now visible: According to (14) with (15) inserted, an economy with twice the propensity to save will have a $2^{\beta/\alpha}$ times higher real wage rate. According to (12) such a real wage rate will induce a $2^{1/\alpha}$ times higher capital intensity. According to (13) such a capital intensity means a twice as high capital co-

efficient. Summing up: The economy with twice the overall propensity to save will have a capital coefficient twice as high. In this sense the adjustment lies on the investment side.

And now for the post-Keynesian model.

III. Equations Peculiar to Post-Keynesian Models

A. *Production*

A post-Keynesian model has fixed input-output coefficients:

$$L = aX \tag{16}$$

$$S = bX \tag{17}$$

Two facts are readily apparent. First, there can be no response of factor proportion to relative factor price, for according to (16) and (17) the factor proportion is $S/L = b/a$, a parameter. Second, (16) and (17) are a simultaneous system implying simultaneous variation of L and S with X. Taking partial derivatives of X with respect to L or S is therefore impossible. Marginal productivities are such partial derivatives. Consequently, to find their distributive shares post-Keynesians need something else than marginal productivities.

B. *Distributive Shares*

The simplest form of a post-Keynesian model has a propensity to consume wages equalling one. In that case the consumption function is

$$C = W/P + c_Z Z/P \tag{18}$$

With immortal capital stock the entire value of output represents value-added, i.e., national money income:

$$Y \equiv PX \tag{19}$$

Insert (4) into (19), divide by P, and write

$$X \equiv W/P + Z/P \tag{20}$$

Subtract (18) from (20) and insert (5). Use (1) and (2) to write $I = g_S S$, insert (17) into that, divide by X, and use (19) to find the profits share:

$$Z/Y = bg_S/(1 - c_Z) \tag{21}$$

C. *Closing the Post-Keynesian Model*

Equation (21) is not a solution yet. So far it merely expresses one unknown, the profits share Z/Y, in terms of another, the rate of growth of capital stock g_S. Do post-Keynesians close their system? At this point Kaldorian and Robinsonian ways are parting.

Kaldor considers the rate of growth of capital stock g_S a variable and solves for it by assuming full employment—as neoclassicists do. Insert (10) into (16), take the derivatives of (16) and (17) with respect to time, use (1), and find that in the absence of technological progress steady-state proportionate rates of growth are

$$g_S = g_X = g_F \tag{15}$$

Insert (15) into (21), and you have a Kaldorian solution for the profits share Z/Y. But to Joan Robinson, g_S is autonomously given by the "animal spirits" of nonprofit-maximizing and otherwise nonrational entrepreneurs. So (21) is already a Robinsonian solution for the profits share.

D. *Does Saving or Investment Adjust to a Higher Propensity to Save?*

If the parametric propensity to save real profits $1 - c_Z$ were twice as high, how would the post-Keynesian model adjust? Let us read equation (21) as follows. If two economies have the same capital coefficient b and are growing at the same proportionate rate g_S, but one economy has a propensity to save real profits $1 - c_Z$ twice as high as that of the other economy, then the former economy will have a profits share Z/Y half that of the latter. That allows the overall propensities to save, and with them the capital coefficients $S/X = b$ to stay the same. In this sense the adjustment lies on the savings side.

E. *The Real Wage Rate, Mark-Up Pricing, and a Possible Wicksell Effect?*

In the post-Keynesian model the real wage rate is hiding behind a pricing formula that in "modern manufacturing industry ... prices are formed by adding a margin to prime cost" (Joan Robinson, p. 179). In a one-good post-Keynesian model "prime cost" is labor cost only, and according to (16) per unit labor cost is aw, hence the formula for price P is $P = amw$ or for the real wage rate

$$w/P = 1/(am) \tag{22}$$

where m is the mark-up factor, and $m > 1$.

Mark-up pricing may be a deviation from neoclassical language but not from neoclassical substance. Under neoclassical pure competition, too, there are overhead costs to be covered, and freedom of entry and exit will see to it that they are, so neoclassical price, too, will exceed "prime cost." The proportion in which it does is easily found. Write (7) as

$$P = \frac{w}{\alpha}\frac{L}{X} \tag{7'}$$

or in English: Price P exceeds per unit labor cost wL/X in the proportion $1/\alpha$. Since $0 < \alpha < 1$, the neoclassical "mark-up" factor $1/\alpha > 1$.

But isn't the post-Keynesian mark-up factor m an interesting new structural parameter reflecting "the degree of monopoly"? If it were, Robinson's system would be overdetermined: Divide (4) by Y, insert (3), (16), (19), and (22), and find another expression for the profits share:

$$Z/Y = 1 - 1/m \tag{23}$$

Consider our two expressions for the profits share (21)—if Kaldorian, with (15) inserted—and (23). If m were a parameter those expressions would be two equations in *one* unknown Z/Y, hence there would be an overdetermined system. If m were a variable those expressions would be two equations in the *two* unknowns Z/Y and m, and we could solve them for m:

$$m = 1/[1 - bg_S/(1 - c_Z)] \tag{24}$$

So we have found the post-Keynesian mark-up factor m to be merely another variable inherent in the savings-investment adjustment mechanism: Robinson's rate of growth of capital stock g_S was given exogenously by the "animal spirits" of entrepreneurs. Thus Robinsonian entrepreneurs may take their pick: Choosing a lower g_S would reduce the profits share (21), would lower the mark-up factor (24), and thereby raise the real wage rate (22).

Do post-Keynesian models have a Wicksell Effect? We might expect none but find one just the same: A higher propensity to save real profits $1 - c_Z$ would reduce the profits share (21), lower the mark-up factor (24), and thereby raise the real wage rate (22). Even a Robinsonian capitalist is "fundamentally the friend of labour."

IV. Conclusions

The present paper has tried to answer four questions: First, is Kaldor correct in saying that in neoclassical models savings determine investment whereas in post-Keynesian models investment determines saving? Taken literally, he is wrong: Savings and investment are both variables determined simultaneously by the parameters of the system. A correctly asked question would be: How sensitive are they to those parameters? What Kaldor really means is that neoclassical and post-Keynesian models have very different sensitivities to the parametric propensity to save. In the neoclassical model, doubling that propensity was found to double the capital coefficient something on the investment side. In the post-Keynesian model, doubling the propensity was found to halve the profits share—something on the savings side.

Second, is the post-Keynesian mark-up factor an interesting new structural parameter reflecting "the degree of monopoly"? Our answer was no: We found it to be merely another variable inherent in the savings-investment adjustment mechanism.

Third, will a Wicksell Effect emerge? We found an expected effect in the neoclassical model and a perhaps unexpected one in the post-Keynesian model.

Fourth, are the models examined open or closed? Yes and no! Kaldor's profit share is determined by his full-employment assumption requiring capital stock, output, and available labor force to be growing at the same rate (15). As James Tobin pointed out so delightfully, the Kaldorian system may be special, arbitrary, and rigid. The burden of adjustment it imposes upon the distributive shares may be heavy, and the distributive shares of the real world may not actually be carrying such a burden; the capital coefficient is more likely to be carrying it. Among the few advanced economies offering usable data on such things, the capital coefficient seems to vary more than do the distributive shares. The United States during 1953–69 had a net propensity to save of 0.081 and a capital coefficient of 2.28. (See the author 1973, p. 35.) Both are roughly one-half of their Swedish counterparts: Sweden, according to Assar Lindbeck (p. 172) has a net propensity to save of 0.14 and according to Erik Lundberg (p. 111) a capital coefficient of 4 to 5, but according to Karl Jungenfelt a labor's share of 0.70 – much like that of the United States.

Still, at least the Kaldorian system is a closed one. By contrast, Robinson's system is an open one. Her profits share as well as her mark-up factor are determined by letting nonprofit-maximizing and otherwise nonrational entrepreneurs fix an arbitrary growth rate of capital stock. Analytical economists consider such openness a deficiency. But perhaps the very openness appeals to interventionists: Control that growth rate and you control income distribution! Perhaps, then, the appeal of post-Keynesian distribution theory is ideological rather than analytical; neoclassical theory is more flexible and has less to fear from confrontation with the real world. (See the author, 1977.)

REFERENCES

Hans Brems, *Labor, Capital, and Growth*, Lexington 1973.

______, "Reality and Neoclassical Theory," *J. Econ. Lit.*, Mar. 1977, *15*, 72–83.

Karl G. Jungenfelt, *Löneandelen och den ekonomiska utvecklingen*, Stockholm 1966.

N. Kaldor, "A Model of Economic Growth," *Econ. J.*, Dec. 1957, *67*, 591–624.

______, "Marginal Productivity and Macroeconomic Theories of Distribution," *Rev. Econ. Stud.*, Oct. 1966, *33*, 309–19.

Assar Lindbeck, *Swedish Economic Policy*, Berkeley; Los Angeles 1972.

Erik Lundberg, *Produktivitet och räntabilitet*, Stockholm 1961.

Joan Robinson, *The Accumulation of Capital*, London 1956.

______, "Solow on the Rate of Return," in G. C. Harcourt and N. F. Laing, eds., *Capital and Growth*, Baltimore 1971, 168–79.

R. M. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Feb. 1956, *70*, 65–94.

J. Tobin, "Towards a General Kaldorian Theory of Distribution," *Rev. Econ. Stud.*, Feb. 1960, *27*, 119–20.

Knut Wicksell, *Föreläsningar i nationalekonomi I*, Lund 1901; trans. as *Lectures on Political Economy I*, London 1934.

The North-South Differential and the Migration of Heterogeneous Labor

By DON BELLANTE*

The differential in median income or earnings between the North and South of the United States has persistently intrigued economists. The consensus view is that: (a) the North-South differential is substantial and has stubbornly persisted through time; and (b) while the labor force responds somewhat to regional pay differentials, the response is not adequate enough to eliminate the differential within a reasonable period of time, since outmigration from low-wage areas is frequently nearly cancelled out by simultaneous inmigration.[1] The consensus view that the labor market is an inefficient allocator of the human resource has been questioned by Philip Coelho and Moheb Ghali in the only study investigating *real* wage differences. When controlling for industry mix at the two-digit level, they found no significant difference between the North and South in average real wages of production workers in manufacturing. However, economists are not prone to discard firmly held conclusions on the basis of a single item of research. For example, a study by Paul Lande and Peter Gordon ignores regional price level variations in finding, at best, very weak support for the neoclassical hypothesis of converging wage levels. Thus, the purpose of this paper is to provide additional evidence on the real North-South differential. Specifically, this paper examines an alternative source of data, the 1970 Census, in order to determine whether more recent and comprehensive data subjected to a different methodology will tend to support or contradict the findings of Coelho and Ghali. In Section I it is argued that analysis of the North-South differential and its migration consequences requires a recognition of the heterogeneous nature of labor. Section II analyzes the statistical findings on the ratio of southern to northern earnings both in the aggregate and in terms of relatively homogeneous subgroups. The migration patterns of the subgroups are also investigated. Conclusions are drawn in Section III. The results are consistent with a neoclassical model of heterogeneous labor.

I. Heterogeneous Labor and the North-South Differential

Implicitly or explicitly, investigations of the North-South differential typically are based on a production function in which various categories of labor are treated as perfect substitutes for one another.[2] A more realistic model would explicitly recognize the heterogeneity of labor by introducing N classes of labor in the production function:

$$(1) \qquad O = f(L_1, L_2, \ldots, L_N, K)$$

*Associate professor of economics, Auburn University. I am indebted to Philip Coelho, George Borts, Mark Jackson, Richard Saba, and Albert Link for helpful remarks. I am particularly appreciative of the extensive suggestions of James Dunlevy and James Long, and of the computational work of Robert Faulk and Betsy Rankin.

[1]For a comprehensive examination of the question of the North-South differential, see Lowell Gallaway. An expansive review of the literature on internal *U.S.* migration, much of which is relevant to the North-South differential, is presented by Michael Greenwood.

[2]While a variety of production functions have been employed in studies of the North-South differential, all treat labor as homogeneous. Typically researchers have recognized the heterogeneity of labor, if at all, by adding *ad hoc* variables to their wage equations. For example Gerald Scully's study of interstate money wage differentials implicitly recognizes the heterogeneity of labor by including per capita state human capital in his industry wage equations. However, the wage data are for production workers in manufacturing by industry and state, whereas the human capital figures are derived from education data for the aggregate state labor force.

such that $(\partial O/\partial L_i)/(\partial O/\partial L_j) \neq 1$, for all $i \neq j$; O is output, K is capital, and each of the N labor inputs L_i represents a different combination of raw labor, formal education, and on-the-job training and experience.

Let us assume that: (a) the production functions for each good are the same in both regions and exhibit constant returns to scale; (b) the patterns of demand are similar in both regions; and (c) the overall capital-labor ratio is the same in both regions. While assumptions (a) and (b) are familiar, assumption (c) is based on Scully's finding that northern and southern capital-labor ratios had converged by the time period of our study. It can be shown that if all three assumptions hold, the pretrade ratio of southern to northern pay for each labor factor will be inversely related to that factor's ratio of relative abundance. Stated otherwise:

$$(2) \qquad \frac{y_i^s}{y_i^n} = g\left(\frac{P_i^s}{P_i^n}\right)$$

where y_i^s is the southern real wage of the ith labor factor; y_i^n is the real wage of the same factor in the North; P_i^s is the quantity of the ith factor in the South, expressed as a percent of the total quantity of labor in the South; and P_i^n is the analogous relative abundance of that factor in the North.[3] In equation (2), $g' < 0$ and $g(1) = 1$, where g' is the derivative of the South-North wage ratio with respect to the South-North ratio of relative abundance. Each region would tend to engage in the export of those goods that utilize more intensively that region's more abundant factors. Under ideal conditions, trade in goods generally would equate the prices of goods between the regions, and payment to the various factors of their marginal product will result in a long-run equilibrium in which $y_i^s/y_i^n = 1$ for each L_i. However, the transport costs of traded goods and the local consumption of untradeable goods will prevent the complete interregional equalization of returns to each L_i. Let us assume that for each factor, interregional trade will remove a fraction of the percentage divergence between northern and southern pretrade returns to that factor. Posttrade ratios of southern to northern factor returns will then be

$$(y_i^s/y_{i\,post}^n) = 1 + X[(y_i^s/y_{i\,pre}^n) - 1]$$

where the subscripts *post* and *pre* identify posttrade and pretrade factor return ratios, respectively, and X is the fraction of the return differential that remains after trade. If X is approximately equal across factors, there will still be a posttrade relation between relative factor abundance and the ratio of southern to northern factor earnings. Further, that relation will still be of the same general form as equation (2), but any given value of the independent variable will correspond to an earnings ratio closer to 1.0 than in the pretrade version of equation (2). Further, we should no longer necessarily expect that $g(1) = 1$.[4]

If for any L_i, posttrade $y_i^s/y_i^n \neq 1$, income-maximizing behavior will result in net migration until that factor's return is equalized between North and South.[5] In terms of a ratio of southern to northern outmigrations, the expected relation is

$$(3) \qquad \frac{M_i^s}{M_i^n} = h\left(\frac{y_i^s}{y_i^n}, A_i\right)$$

where M_i^s and M_i^n are gross outmigration of the ith labor factor from the South and North, respectively. A ratio of M_i^s/M_i^n greater than one indicates net outmigra-

[3]While reliance on Scully's empirical observation may be theoretically inelegant, it makes possible the avoidance of some awkward mathematics without a corresponding loss of substance.

[4]The inability of trade to equalize prices between regions is amply demonstrated by the data on comparative living costs which are discussed in the next section. Further, while there is no a priori reason to expect X to be approximately the same for all categories of labor, the empirical evidence of the next section appears to justify the assumption.

[5]While nonzero migration occurs in response to a real earnings differential, it should be pointed out that other factors will tend to move toward the region of highest nominal return, since the owners of those factors need not live in the receiving region.

tion from the South for the ith labor factor. The term A_i is the age of the ith labor factor.[6] Both the posttrade version of the short-run relative wage relation of equation (2) and the long-run adjustment process expressed in equation (3) can be tested with existing cross-sectional data.

The discussion of this section has some implications for empirical examinations of the North-South differential. The most obvious implication is that a simple average differential of wages or earnings is inappropriate. Meaningful statements about the average North-South wage or income ratio require construction of an index of wages or income which averages the same quantities, both in the North and in the South, of each distinguishable labor factor. Secondly, the apparently paradoxical observation that aggregate gross inmigration to and outmigration from an area are highly correlated can be readily explained. An implication of my heterogeneous labor model is that an overabundance of one class of labor implies that one or more other classes are undersupplied. Thus, cross movements of the labor force do not frustrate the efficient allocation of labor, as is frequently alleged. On the contrary, the tendency of high inmigration areas also to be high outmigration areas is to be expected and indicates a higher degree of allocative efficiency than would be indicated by a predominately unidirectional movement of labor. In the next section, the available evidence is examined with respect to both of these implications. As a by-product of the statistical analysis, we are able to identify the sources of the differences in the unweighted averages of northern and southern earnings levels.

II. Statistical Findings

A proper index of the ratio of southern to northern earnings would be a weighted average of the relative earnings of each of the N categories of labor. For a Paasche-type index, the weight of the ith category would correspond to the percent of the South's total workforce made up by the ith category: $\Sigma y_i^s P_i^s / \Sigma y_i^n P_i^s$. Conversely, a Laspeyeres-type index would use northern weights: $\Sigma y_i^s P_i^n / \Sigma y_i^n P_i^n$. Since the choice of index type is for the most part arbitrary, the methodological discussion will be limited to the Paasche Index. Nonetheless, the two methods will yield somewhat different results; hence empirical results will include both measures.

While the desirability on theoretical grounds of distinguishing between various labor inputs on the basis of formal education and on-the-job training and experience is apparent, there are problems in operationally identifying the various classes. The traditional procedure in identifying earnings functions has been to treat earnings as a function of years of schooling, age, and race, with age and race together serving as proxies for experience and on-the-job training. Age is a direct proxy for experience and is also a determinant of the cumulative past availability of investments in on-the-job training. Race takes into account the effects of discrimination on the availability of on-the-job training and the availability of "meaningful" work experience. Since the tendency to invest in on-the-job training is known to be positively associated with years of formal education, the educational variable will also serve as a partial proxy for on-the-job training. The variables of schooling, age, and race can thus be used, as they will be below, as bases for categorizing labor into relatively homogeneous groups. However, as Jacob Mincer and Solomon Polachek have pointed out, the proxy relationship between age and experience and on-the-job training holds only for workers with reasonably continuous work histories and not for those workers who expect to withdraw periodically from the labor force. For that reason my analysis is limited to the so-called primary labor force—males ages 25-64. Workers in this group have a very high and consistent rate of labor force participation.

[6]Age is one of the characteristics that distinguishes one labor factor from another, acting as a proxy for experience and specificity of skills as will be explained below. As is well known, age will negatively affect the migration response to a given regional earnings differential.

The 1970 Census reports 1969 regional per capita earnings of males cross classified by years of schooling, age, and race. Within the primary labor force, the earnings data are cross classified into four age groups (25-34, 35-44, 45-54, 55-64), eight educational attainment groups (less than 5 years, 5-7 years, 8 years, 1-3 years of high school, 4 years of high school, 1-3 years of college, 4 years of college, 5 or more years of college), and two racial groups (white and nonwhite). There are thus sixty-four age-education-race categories in each of the two regions.

In order to compare earnings in real terms, southern earnings must be inflated to reflect the lower cost of living in the South. First an index of living costs in the North was constructed as $C_n = \Sigma C_i w_i / \Sigma w_i$ where C_i is the cost-of-living index for northern city i, and w_i is the weight that each northern city is given by the Bureau of Labor Statistics (*BLS*) in their construction of the *U.S.* average cost of living. The C_i and w_i are from the spring 1969 calculations of the intermediate level budget. For the North, *BLS* used thirty Standard Metropolitan Statistical Areas (*SMSA*s) in addition to nonmetropolitan areas. A similar index C_s was constructed for the South, which includes nine *SMSA*s plus the surveyed nonmetropolitan areas.[7] Next an inflator was formed, $I = C_n/C_s$. Each southern income cell is then multiplied by I. The complex weighting and sampling procedure employed by *BLS* assures that my inflator will be a reasonable measure of comparative living costs.[8]

From the real earnings data, a Paasche Index can be constructed which in effect shows what the South-North real earnings ratio would be if the South had the same distribution of educational-demographic characteristics as the North. In other words, the ratio is an index of relative earnings, using the same weights for each of the sixty-four classes of labor in the North as are used for the South. The ratio, which will be referred to as the "Indexed Real Earnings Ratio," is defined as

$$(4)\quad \sum_a \sum_e \sum_r y^s_{aer} P^s_{aer} \Big/ \sum_a \sum_e \sum_r y^n_{aer} P^s_{aer}$$

where y^s_{aer} is the mean real earnings in the ath age, eth educational, and rth racial category in the South; y^n_{aer} is similarly defined for the North; and P^s_{aer} is the population in each *aer* category expressed as a percent of the total of all sixty-four southern categories.

Table 1 presents the indexed real earnings ratio and, for comparison an "Indexed Money Earnings Ratio" that is similar to the indexed real earnings ratio but does not use cost-of-living inflated earnings data. Table 1 also contains a "Demographic Ratio" defined as

$$(5)\quad \sum_a \sum_e \sum_r y^n_{aer} P^s_{aer} \Big/ \sum_a \sum_e \sum_r y^n_{aer} P^n_{aer}$$

The demographic ratio measures the extent to which the age, educational, and racial differences between the labor forces of the South and North result in a difference in mean earnings. In other words the demographic ratio indicates what the South-North earnings ratio would be if the South differed from the North only in terms of its distribution of educational-demographic characteristics and not in terms of the mean earnings of each of the sixty-four classes of labor. The product of the indexed real earnings ratio and the demographic ratio is identically equal to the "Unindexed Real

TABLE 1—SOUTH-NORTH EARNINGS RATIOS

Unindexed Money Earnings Ratio	.8192
Unindexed Real Earnings Ratio	.9330
Indexed Money Earnings Ratio (Paasche method)	.8820
Indexed Real Earnings Ratio (Paasche method)	1.0040
Demographic Ratio	.9288
Age Ratio	.9987
Education Ratio	.9458
Race Ratio	.9695

[7]The "South" as used in this paper is the U.S. Census South. The "North" is the rest of the United States which the Census refers to as the North and West

[8]The intermediate level budget is the budget most often used in studies of comparative living costs. Results using the alternate measures provided by *BLS*, the lower and higher budget levels, will also be reported below.

Earnings Ratio," that is, the simple ratio of southern to northern mean earnings:

$$(6)\quad \frac{\sum_a\sum_e\sum_r y^s_{aer}P^s_{aer}}{\sum_a\sum_e\sum_r y^n_{aer}P^s_{aer}} \cdot \frac{\sum_a\sum_e\sum_r y^n_{aer}P^s_{aer}}{\sum_a\sum_e\sum_r y^n_{aer}P^n_{aer}} \equiv \frac{\sum_a\sum_e\sum_r y^s_{aer}P^s_{aer}}{\sum_a\sum_e\sum_r y^n_{aer}P^n_{aer}}$$

It is also true that the product of the indexed money earnings ratio and the demographic ratio is identically equal to the unindexed money earnings ratio, which is the simple ratio of southern to northern mean earnings, unadjusted for cost-of-living differences.

The three effects (age, education, race) considered in the demographic ratio can be separated by means of further manipulation of the basic data. For example, an age ratio that measures the effect of the differential age distributions on the income ratio can be defined as

$$(7)\quad \sum_a y^n_{a\cdot} P^s_{a\cdot} \Big/ \sum_a\sum_e\sum_r y^n_{aer}P^n_{aer}$$

where $\Sigma_a y^n_{a\cdot} P^s_{a\cdot} = \Sigma_a[\Sigma_e\Sigma_r y^n_{aer}\cdot\Sigma_e\Sigma_r P^s_{aer}]$. The age ratio in effect measures what the South-North earnings ratio would be if: (a) the South maintained its own age distribution but within each age cell had the same racial and educational distribution as the North, and (b) the South had the same mean earnings within each age-education-race cell as the North. Similarly, an education ratio, which measures the effect of the difference in educational attainment between North and South, can be calculated by substituting $\Sigma_e y^n_{\cdot e} P^s_{\cdot e}$ for the numerator in equation (7). In like fashion a race ratio, which measures the effect of the South's greater proportion of nonwhites on the South-North earnings ratio, can be constructed by substituting $\Sigma_r y^n_{\cdot r} P^s_{\cdot r}$ for the numerator of equation (7). The age ratio, and race ratio are also presented in Table 1.

The results of Table 1 indicate that while the simple ratio of southern to northern mean money income is .8192, adjustment for cost-of-living differences raises the simple ratio of southern to northern mean real earnings to .933. After removing the effects of differences between the North and South in their price levels and their age, educational, and racial distributions, the indexed real earnings ratio is very close to 1.0.

The value of the demographic ratio indicates that differences in the mix of labor endowments between North and South alone would result in real southern earnings being about 93 percent of what they are in the North. The age ratio of nearly 1.0 would seem to indicate that there is virtually no difference between the age structure of the northern and southern labor forces, but such is not the case: The South has a disproportionately lower share of older workers (ages 54–65) and a greater share of younger workers (25–34) than the North. However, both the under- and overrepresented age groups have mean earnings below those of the two middle-age groups, with the result that the effects of both groups largely cancel out. On the other hand the calculation of the education ratio demonstrates that if the only factor that differed between North and South was educational attainment, average real earnings in the South would be only about 95 percent of northern real earnings. Further, the value of the race ratio indicates that the disproportinately high representation of nonwhites in the South would by itself result in southern mean earnings being about 3 percent below the northern mean. The use of northern earnings levels for the calculation of the race ratio does not take into account regional differences in wage discrimination against nonwhites.[9]

The near unity value of the indexed real earnings ratio thus supports the finding of Coelho and Ghali that the North-South differential does not exist. The differences

[9]When the race ratio is reconstructed using southern rather than northern earnings levels—and thus, implicitly, southern magnitudes of racial discrimination—its value becomes .9569. The effect of discrimination on earnings of nonwhites may be largely offset by resultant increases in white earnings, leaving the aggregate calculations of this paper only modestly affected. In any event, a full discussion of the effects of racial discrimination is quite complex and beyond the scope of this paper.

between the present study and theirs are then noteworthy. Coelho and Ghali base their finding of no differential on the fact that in regression of real wages on a series of independent variables, a dummy variable representing the South was not significantly different from zero. The variables included 2-digit industry classification, sex, color, capital-labor ratio, and average educational attainment of each metropolitan area. In perfectly competitive wage theory, no industrial differentials should exist except to the extent that industries differ in their required human capital mixes. If that were the only effect measured by their industry dummies, the Coelho and Ghali study would in principle be quite similar to the present study. In such a case, their use of some other human capital oriented variables, particularly educational attainment, would have been unnecessary but would not have biased their regional dummy.[10] Indeed, inclusion of the variable does not materially alter their finding. Yet, potential critics of the Coelho and Ghali approach might argue that much more than differential human capital requirements are captured in their industry dummies. For instance, the industries for which wages are most out of line with competitive levels, because of unionization or other effects, could be disproportionately concentrated in the North. If so, the coefficient on their regional dummy variable would not reflect the full regional differential. The use of the index method in the present study avoids the potential problem of colinearity between a regional variable and other independent variables.

Coelho and Ghali examined earnings in manufacturing in large *SMSA*s, thus they excluded a large segment of the labor force. Given the degree of unionization in manufacturing, the North-South differential might be expected to be particularly small in this sector. The present study includes the entire labor force of the South. It is not being argued that the methodology of this study is superior to that of Coelho and Ghali. Indeed, my methodology creates problems of its own such as the previously alluded to index number problem. The purpose of this discussion is to emphasize the point that the present study's methodology is substantially different from that of Coelho and Ghali, yet the conclusions are the same.[11]

A. *Alternative Specifications*

Given that the Paasche ratio of southern to northern earnings, using an intermediate budget deflator, tends to support the Coelho-Ghali finding of no significant North-South differential, we examine whether alternative specifications will contradict the conclusion. Specifically, the South-North ratio is calculated using all three budget levels as inflators and using both Paasche and Laspeyeres indexes of real earnings. Thus far, the comparison has been between the South and the entire North, or non-South. In what follows, the South is compared not only with the North, but with the three northern subregions: the Northeast, North Central, and West. Table 2 presents all twenty-four combinations of region, inflator, and index method.

The ratio of southern to northern earnings ranges from .965 to 1.025, depending on which cost-of-living inflator and which index method is used. In no case is there evidence in support of a substantial deficiency in real earnings in the South. The earnings ratios are more varied when the South is compared to the three subregions of the North. The ratios in these cases are

[10]The possibility of multicollinearity between their education and industry variables may explain the wrong sign obtained on their education variable, since a wrong sign is usually obtained, as between two collinear independent variables, for the one having the lower simple correlation coefficient with the dependent variable.

[11]Since most studies of the North-South differential employ regression methods, I have also employed the regression technique. When real earnings are regressed on a series of dummy variables representing the age, education, and race categories, and a dummy variable representing the South, the coefficient of the South variable is insignificant and indicates a South-North real earnings ratio of about .98. However, since each category was weighted equally, the regression result is not comparable to our weighted index technique.

TABLE 2 - REAL EARNINGS RATIOS

Earnings Ratio	Index Method	Budget Level used for Cost-of-Living Inflator: Lower	Intermediate	Higher
South to Northeast	Paasche	.955	1.033	1.063
	Laspeyeres	.938	1.002	967
South to Northcentral	Paasche	.978	1.004	1.009
	Laspeyeres	1.001	.983	.996
South to West	Paasche	1 072	1.087	1 058
	Laspeyeres	1.060	1.042	1.057
South to North	Paasche	.965	1 004	1 025
	Laspeyeres	986	1.028	1.012

perhaps less reliable since the sample of cities in each subsample is small. In any event, only the Paasche ratio of southern to northeastern real earnings using the lower budget deflator is substantially less than 1.0. Coelho and Ghali (1973, p. 759) argued against the use of the lower budget, as well as the higher budget, in their response to criticism by Mark Ladenson. Whether or not results using the lower budget should be discounted, the fact remains that the differential of 6.2 percent is about half the size of the lowest Northeast-South differential reported in studies prior to Coelho and Ghali.

In all comparisons using intermediate or higher budget inflators, southern real earnings are lower by no more than 3.2 percent. In most cases, southern real earnings are higher than the region or subregion of comparison. Somewhat surprising is the finding that regardless of which index method or deflator is used, real earnings are at least 4.2 percent higher in the South than in the West. On balance, the results of Table 2 provide additional evidence against a deficiency of southern wage levels. When the South is compared to the entire North, the ratio of southern to northern real earnings is in all cases quite close to 1.00.

B. *Migration of Labor between North and South*

Although a properly weighted average of real earnings has revealed no overall North-South differential, it is by no means the case that the real earnings ratio is close to 1.00 for each of the sixty-four age-education-race groups. The actual ratios ranged from .86 to 1.11. Our discussion in Section I leads to the prediction, stated in equation (2), that in short-run equilibrium the ratio of southern to northern earnings for each age-education-race group will be inversely related to that group's South-North ratio of relative abundance. In fact, simple linear ordinary least squares (*OLS*) regression of the sixty-four earnings ratios on their corresponding ratios of relative supply yields the following empirical version of equation (2), with the t-ratio in parenthesis:

$$\text{(8)} \qquad \frac{y^s_{aer}}{y^n_{aer}} = \underset{}{1.036} - \underset{(10.72)}{.057}\left(\frac{P^s_{aer}}{P^n_{aer}}\right)$$

The R^2 for equation (8) is .65 and the t-ratio is significant at the .0001 level. The observed earnings ratios are therefore consistent with the short-run, posttrade predictions of competitive theory. Interestingly, for $P^s_{aer}/P^n_{aer} = 1$, the predicted value of y^s_{aer}/y^n_{aer} is .979. This value is close to (and not significantly different from) the pretrade value of 1.00 that we would expect to observe if in fact pretrade earnings ratios were observable.[12]

The long-run adjustment mechanism described in equation (3) suggests that the ratio of southern to northern outmigration

[12]The finding of nonsignificance is based on a t-ratio of .076. The relevant t-ratio is derived in John Johnston, pp. 42-43

(stated otherwise, the ratio of southern outmigration to southern inmigration) for each age-education-race group will be inversely related to that group's age. The available data permit us to examine the predicted relations.[13] The empirical version of equation (3) must take into account the expectation that for any given earnings ratio, the speed of response must depend on the age level of each observation. The speed of response to an earnings differential should decline with age. For this purpose, a series of dummy variables is used. The youngest age group (25–34) is the basis group. A dummy variable A_1 represents the 35–44 age group, A_2 represents the 45–54 age group, and A_3 designates the 55–64 age group. The empirical version of equation (3) yields the following linear *OLS* result, with *t*-ratios in parentheses:

$$(9)\quad \frac{M^{s*}}{M^{n*}} = -.11 \underset{(-8.76)}{-8.00}\frac{y^{s*}}{y^{n*}} + \underset{(2.70)}{3.33}A_1\cdot\frac{y^{s*}}{y^{n*}} + \underset{(4.71)}{5.57}A_2\cdot\frac{y^{s*}}{y^{n*}} + \underset{(6.61)}{7.58}A_3\cdot\frac{y^{s*}}{y^{n*}}$$

where $M^{s*}/M^{n*} = M^{s}_{aer}/M^{n}_{aer} - 1$ and $y^{s*}/y^{n*} \equiv y^{s}_{aer}/y^{n}_{aer} - .1$ The value of R^2 for equation (10) is .70 and $F = 24.32$. All of the variables are significant at the .0001 level, except the coefficient of $A_1 \cdot y^{s*}/y^{n*}$, which is significant at the .01 level. Further, not only the signs but also the relative magnitudes of the coefficients of the terms involving the age dummies are as predicted.[14] As would be expected, the intercept is not significantly different from zero at the .10 level. The results indicate that for the 25–34 age group, for each percentage point by which southern earnings are below the northern level, there will be an 8 percent excess of southern outmigration over inmigration. The elasticity for the 35–44 age group is 4.67, for the 45–54 age group it is 2.43, and for the 55–64 age group, the elasticity is .42. Since the migration data employed in this study are for a five-year period (1965–70), the indicated response to a given differential should be interpreted as a five-year rate of adjustment, not an annual rate.

The coefficients revealed in equation (9) should not be taken as precise estimates. The true relationship between migration and earnings differentials is of course simultaneous. The 1969 earnings data come very close to being end-of-period earnings: 1965 earnings data, though not available, would have been preferred. However, given that migration should narrow the differential, the direction of bias is such that 1965 earnings data should have yielded greater elasticities. In any event, the conservative elasticity estimates of equation (9) suggest a rather strong response to the existence of a real earnings differential.

III. Summary and Implications

The study of the North-South differential without adjustment for cost-of-living differences can be misleading, since over 62 percent of the difference between southern and northern per capita money earnings is attributable to differences in living costs, as measured by *BLS*'s intermediate level budget. The remaining 38 percent can be explained by differences between the North and South in the mix of human capital-related demographic characteristics. When the effect of the differential mixes is removed, the data from the 1970 Census support the earlier finding of Coelho and Ghali that the North-South differential does not exist. The finding is most strongly supported when the Paasche Index method is applied to real wages calculated with an

[13]The migration data are derived from the U.S. Census subject report *Lifetime and Recent Migration* The age, race, and education categories of the migration data are the same as for the earnings data used above, except that the highest and lowest educational categories of the migration data covered the two highest and two lowest educational categories of the earnings data.

[14]Since the relation between the outmigration ratio and the earnings ratio is inverse, an expected negative relation between the outmigration ratio and age would be demonstrated in equation (9) if $|g'(y^s/y^n)| > g'(A_3) > g'(A_2) > g'(A_1) > 0$.

intermediate level cost-of-living inflator. Once account is taken of the heterogeneity of labor, the observed patterns of real earnings and migration are consistent with purely competitive theory. The response of migration is much stronger than has been indicated by studies that aggregate over the various categories of human capital, particularly educational categories.

Larry Sjaastad long ago suggested that the apparent inefficiency of migration may simply be the result of aggregation. The fact that few studies have disaggregated labor to the extent that we have may account for Michael Greenwood's statement, in his survey of the literature, that "[J. R.] Hicks' contention that wage differences are the chief determinant of migration has not been confirmed" (p. 411). Yet at the level of disaggregation employed in equation (9), 70 percent of the variation in the ratio of southern outmigration to inmigration can be explained by the age-adjusted ratio of southern to northern earnings alone.

The finding of a South-North real earnings ratio close to 1.0 might be made more precise by further manipulating the data. For example, the division into sixty-four classes of labor is arbitrarily determined by the format in which the Census data are reported, and the earnings data are not corrected for regional differences in hours of work. However, further manipulations seem unlikely to result in a ratio greatly different from unity.

Given the limitations in available data, neither this study nor the Coelho and Ghali study can be accepted as final proof of the efficiency of the market mechanism with regard to the allocation of labor between North and South. It may well be that real earnings have always been significantly lower in the South and have risen to rough equality with the North only since the late 1960's. For that matter, real wages in the South may prove to have risen above northern levels by the time of the upcoming 1980 Census, as some parties to the Sunbelt-Snowbelt controversy seem to fear. If so, the apparent support for the simple, static version of the neoclassical model may be merely the result of having used data gathered at the point in time when real wages in the North and South happened to cross. Yet because we are without reliable historical regional price data, we are almost as uncertain about what magnitudes of real differentials have existed in the past as about the size of future differentials. However, since those economists who have found fault with neoclassical models are prone to make recommendations that involve substantial intervention into the realm of private decision making,[15] it seems appropriate that for purposes of present and future policy recommendations, those scholars now should bear the burden of proof.

[15]As an extreme example, Rufus Hughes has called for government intervention to obtain: 1) a more equal geographical distribution of investment, 2) restrictions on interregional trade, 3) more educational and vocational preparation programs for underdeveloped regions, and 4) mobility patterns favorable to the underdeveloped regions of the South.

REFERENCES

P. R. P. Coelho and M. A. Ghali, "The End of the North-South Wage Differential," *Amer. Econ. Rev.*, Dec. 1971, *61*, 932–37.

——— **and** ———, "The End of the North-South Wage Differential: Reply," *Amer. Econ. Rev.*, Sept. 1973, *63*, 757–62.

L. E. Gallaway, "The North-South Wage Differential," *Rev. Econ. Statist.*, Aug. 1963, *45*, 264–72.

M. J. Greenwood, "Research on Internal Migration in the United States: A Survey," *J. Econ. Lit.*, June 1975, *13*, 397–433.

R. B. Hughes, "Interregional Income Differences: Self-Perpetuation," *Southern Econ. J.*, July 1961, *58*, 41–45.

John Johnston, *Econometric Methods*, New York 1972.

M. L. Ladenson, "The End of the North-South Wage Differential: Comment," *Amer. Econ. Rev.*, Sept. 1973, *63*, 754–56.

P. Lande and P. Gordon, "Regional Growth in the United States: A Reexamination of the Neoclassical Model," *J. Reg. Sci.*, Apr. 1976, *17*, 61–69.

J. **Mincer and S. Polachek**, "Family Investments in Human Capital: Earnings of Women," *J. Polit. Econ.*, Apr. 1974, *82*, S76-S108.

G. W. **Scully**, "Interstate Wage Differentials: A Cross Sectional Analysis," *Amer. Econ. Rev.*, Dec. 1969, *59*, 757-73.

L. A. **Sjaastad**, "The Costs and Returns of Human Migration," *J. Polit. Econ.*, Oct. 1962, *70*, S80-S93.

U.S. Bureau of the Census, *U.S. Census of Population: 1970; Vol. I, Characteristics of the Population*, Part 1, U.S. Summary, sec. 2, Washington 1973.

———, *U.S. Census of Population: 1970, Subject Reports, Lifetime and Recent Migration*, PC(2)-2D, Washington 1973.

U.S. Bureau of Labor Statistics, *Three Budgets of Living for an Urban Family of Four Persons*, Bull. 1570-5, Washington 1970.

Short- and Long-Run Effects of Monetary and Fiscal Policies under Flexible Exchange Rates and Perfect Capital Mobility

By CARLOS ALFREDO RODRIGUEZ*

Following the pioneering papers by John M. Fleming and Robert Mundell, a substantial literature has accumulated incorporating the Keynesian analysis of the effects of monetary and fiscal policies in the open economy under flexible exchange rates and perfect capital mobility. The general policy result which followed from this line of research has been the verification of the presumption that monetary policy is an effective stabilization tool under flexible exchange rates, while the ability of fiscal policy to affect the level of economic activity varies inversely with the degree of international capital mobility. The key to the results lies in the differential effect of both policies on the direction of the induced capital flows and thus on exchange rate movements. The latter in turn affect the trade balance and aggregate demand. To my knowledge, however, all authors have either been concerned with the derivation of short-run or "impact" multipliers describing the effects of monetary and fiscal policy on the level of economic activity or have otherwise ignored the "longer-run" effects of policy-induced changes in the level of international indebtedness and the *service* of that debt.

Abstracting from growth and persistent shocks, it is reasonable to assume that, given enough time following the policy change, individuals will adjust their international portfolios of assets to the new desired proportions such that capital flows eventually cease. When long-run portfolio equilibrium is thus reached, the service account deficit of the balance of payments must necessarily be equal to the net export surplus (the trade balance), the exchange rate being the natural instrument through which this long-run external balance condition is achieved. It is therefore natural to conceive of the long-run level of the service account as the primary determinant of the long-run trade balance. In view of the above and the fact that in a Keynesian-type economy the trade balance plays a crucial role in the determination of the level of economic activity through its effects on aggregate demand, it follows that a longer-run analysis of the effects of monetary and fiscal policy cannot logically ignore the consequences of those policies for the service account.

In this paper I intend to explore the long-run implications of induced changes in the level of the service account for the effects of monetary and fiscal policy; in the process of doing so, I will develop a simple model whose basic structure resembles that of Fleming and Mundell although it is modified to explicitly incorporate the stocks of domestic and foreign securities held.

In order to provide the reader with a clearer perspective on the problem at hand, let us first discuss the impact effects of monetary and fiscal policy in the context of the standard short-run model. Given the Keynesian structure of the economy either policy will increase domestic income (and employment) only to the extent that it succeeds in expanding aggregate demand of which the trade balance is one component. At a given exchange rate, expansionary fiscal policy would increase aggregate demand but at the expense of a higher domestic interest rate (the "crowding out" effect);

*Columbia University. I am indebted to G. Borts, L. Girton, and D. Henderson for their comments and suggestions. A preliminary version of this paper was written while I was visiting scholar at the Division of International Finance of the Board of Governors of the Federal Reserve System. The views expressed here should not be interpreted as necessarily those of the Board.

however, the assumption of perfect international capital mobility prevents the interest rate from rising. In response to the upward pressure on the interest rate capital flows in, the currency appreciates, and as a consequence, the trade balance deteriorates, offsetting the expansionary effects of the fiscal stimulus on aggregate demand. Only when the trade balance has deteriorated enough to bring aggregate demand to the prefiscal expansion level will the interest rate be the same as abroad and short-run equilibrium thus be restored. Therefore, a fiscal expansion is nullified by an induced deterioration of the trade balance triggered by a capital inflow.

Expansionary monetary policy, on the other hand, increases aggregate demand through a lower interest rate (at a given exchange rate). The resulting capital outflow depreciates the currency, improves the trade balance, and thus contributes further to the expansion of demand. The higher income due to the depreciation raises money demand such that the interest rate is brought back to the initial level and equilibrium is restored. It follows that the induced short-run capital outflow further compounds the expansionary effects of monetary policy.

Notice that since a capital outflow is equivalent to a positive rate of acquisition of foreign securities the above argument implies that the net foreign asset position of the country tends to improve with a monetary expansion and, conversely, to deteriorate with a fiscal expansion. In the long run, if the system is stable, portfolio holders will be satisfied with the level and composition of their assets, and capital flows will cease. At that point, given the world's interest rate, the service account is fully determined by the level of the net asset position of the country. In the case of a fiscal expansion, the long-run effect of a transitional period of induced capital inflows is therefore to deteriorate the long-run service account while the induced capital outflows due to expansionary monetary policy would improve it. Since in the long run the capital account will be zero, the service account deficit must equal the trade account surplus and therefore it follows that, on account of the induced capital flows, the long-run effect of expansionary fiscal policy is to improve the trade balance while expansionary monetary policy deteriorates the trade balance in the long run. Notice that the long-run effects of both policies on the trade balance are precisely the *opposite* from those in the short run.

Since the trade balance is one of the determinants of aggregate demand, the question arises as to whether in the long run a monetary expansion, which deteriorates the trade balance, may not actually induce a *fall* in income and employment; similarly, to the extent that expansionary fiscal policy works towards an improved long-run trade balance, it may be able to increase income in the long run in spite of its short-run ineffectiveness. In order to formally verify the above conjectures the next section develops a simple dynamic model which incorporates the actual stocks of securities held and uses national income instead of domestic income as the relevant variable in all behavioral relations (the two concepts differ by the amount of net interest payments abroad).

I. The Model

To simplify the presentation, I will assume that there is no net creation of domestic securities (no domestic investment) and thus the only components of aggregate demand are domestic consumption C, government expenditures G, and the trade balance surplus T. Domestic output (equal to domestic income) is in perfectly elastic supply at a constant money price (assumed equal to unity). Goods market equilibrium requires

$$Y = C + G + T \tag{1}$$

where Y is domestic income. There are three types of assets in this economy:

a) Domestic money M held only by domestic residents and supplied by the central bank.

b) Government bonds, each paying one unit of domestic currency per unit of time in perpetuity. Assuming the domestic inter-

est rate is expected to remain constant, the price of each bond is the inverse of the interest rate. Denoting the total flow of payments on government bonds and the interest rate by B and r, the total market value of the stock is then B/r. Government bonds are assumed not to be internationally traded.

c) An internationally traded bond, issued abroad and paying one unit of foreign exchange per unit of time in perpetuity. The interest rate abroad is fixed at r_o and thus the price of the bond in terms of foreign exchange is $1/r_o$. The total stock of foreign bonds held by domestic residents is denoted by D, which is also the value of interest payments. The value in terms of domestic currency of the stock of foreign bonds held in the country is then eD/r_o, where e is the exchange rate. Domestic and foreign bonds are assumed to be considered perfect substitutes by portfolio holders and thus the interest rate on government bonds, assuming static expectations with respect to future levels of the exchange rate, must be equal to r_o.

Monetary policy is defined as a once and for all exchange between money and government bonds by the central bank ($\Delta M = -\Delta B/r_o$). Neither money nor bonds are currently issued to finance operations of the treasury; thus, current treasury spending G plus the service of the outstanding government debt outside the central bank B must equal taxes Z:

$$G + B = Z \tag{2}$$

The outstanding stock of government bonds is therefore the result of accumulated past deficits of the treasury and no new issues of such bonds take place. While it would be possible to interpret fiscal policy in terms of deficit spending (financed through new bond issues) such a policy, in the absence of growth, is inconsistent with the existence of a steady state since the stock of bonds would tend to increase without bounds. I thus prefer to consider only the case when fiscal policy takes the form of a balanced increase in both government spending and taxes, as required by equation (2).

National income is domestic output plus interest payments from foreigners, $Y + eD$; notice that we need not restrict D to be positive, meaning that the country may be a net debtor to the rest of the world but her indebtedness must be denominated in terms of foreign currency (allowing for the ownership of external assets or indebtedness to be denominated in terms of domestic currency does not change any of the qualitative results of the model and reinforces its stability properties).

Disposable national income Y^d equals national income plus interest payments from the government minus taxes: $Y^d = Y + eD + B - Z$. Given the treasury's budget constraint (2), Y^d becomes

$$Y^d = Y + eD - G \tag{3}$$

Consumption is assumed to depend positively on Y^d:

$$C = C(Y + eD - G) \qquad 0 < C' < 1 \tag{4}$$

The demand for domestic money depends positively on both the value of other assets held and national income:[1]

$$M^d = L(Y + eD, (eD + B)/r_o) \qquad L_1 > 0, L_2 > 0 \tag{5}$$

The trade balance depends negatively on Y^d and positively on the exchange rate (I assume that the standard Marshall-Lerner condition is satisfied):

[1] It is debatable whether national income or disposable income should be the correct transactions variable in the demand for money. The interested reader can easily verify that when Y^d is used none of the results obtained here for monetary policy are changed while those obtained for fiscal policy remain qualitatively the same in the long run although they differ in the short run: the short-run output multiplier for a balanced increase in government expenditures and taxes becomes unity rather than zero as in Fleming-Mundell. The assumption that money demand depends on income in addition to assets, although ignored in most of the recent "portfolio balance" literature, has a long standing in the literature on monetary theory: see for example the models developed by Don Patinkin and by Duncan Foley and Miguel Sidrauski. In the context of this model the dependence of money demand on both income and assets is crucial for both the existence and the stability of a long-run solution; this point is discussed further in fn. 3.

$$T = T(Y + eD - G, e) \qquad T_1 < 0,\ T_2 > 0 \tag{6}$$

In what follows it will be assumed that $0 < C' + T_1 < 1$, which amounts to the assumption that the marginal propensity to spend on domestic goods is positive but less than unity.

Finally, the rate of acquisition of foreign securities is equal to the current account surplus:

$$e\dot{D}/r_o = T(Y + eD - G, e) + eD \tag{7}$$

Given the levels of the money supply M, government spending G, government debt B, and the stock of foreign securities held D, the short-run equilibrium of the economy is described by the following two conditions:

Money Market Equilibrium

$$M = L(Y + eD, (eD + B)/r_o \tag{8}$$

Goods Market Equilibrium

$$Y = C(Y + eD - G) + G + T(Y + eD - G, e) \tag{9}$$

Equations (8) and (9) jointly determine the values of domestic income Y and the exchange rate e as functions of the government determined parameters M, G, and B, and the existing stock of foreign securities D. Equation (7), in turn, determines the rate at which foreign securities are being acquired from abroad.

Differentiating totally (8) and (9) we obtain the relationship between Y, e, and all predetermined variables:

$$L_1 dY + (L_1 + L_2/r_o)D \cdot de = -e(L_1 + L_2/r_o)dD + dM - (L_2/r_o)dB \tag{10a}$$

$$-(1 - C' - T_1)dY + (DC' + DT_1 + T_2)de = -e(C' + T_1)dD - (1 - C' - T_1)dG \tag{10b}$$

It follows from (10a,b) that holding constant the policy variables M, G, and B, domestic income and the exchange rate change with D according to

$$dY/dD = -eT_2(L_1 + L_2/r_o)/A \tag{11}$$

$$de/dD = -e[L_1 + (1 - C' - T_1) \cdot L_2/r_o]/A \tag{12}$$

where

$$A = L_1T_2 + D[L_1 + (1 - C' - T_1) \cdot L_2/r_o] \tag{13}$$

In order to be able to analyze the long-run comparative static properties of the model, I will assume the existence of a stable steady state. Existence of a steady state implies that eventually the stock of foreign securities will reach a value for which its rate of change is zero, that is, the current account will be zero. Such an equilibrium will be locally stable if following a small increase (decrease) in D, its rate of change becomes negative (positive) such that D converges again to the former value. Algebraically, the steady state is locally stable if the derivative of $\dot{D}$ with respect to D, evaluated around the steady state, is negative. Differentiating (7) and substituting (11) and (12) for the changes in Y and e due to changes in D, we obtain

$$d\dot{D}/dD = -L_2T_2(1 - C')/A$$

which will be negative provided A is positive (which will be guaranteed if the country is a net creditor such that D in (13) is positive). In what follows it will be assumed that A is positive and thus that the system is stable.

II. The Short-Run Multipliers

The short-run effects on income and the exchange rate of monetary and fiscal policies can be readily obtained from (10a,b) while holding $dD = 0$. In the case of monetary policy we assume that the money supply is increased via an open market purchase of government bonds by the central bank ($dM = -dB/r_o$). Notice that making dG positive while holding B and M constant implies that the extra government spending is financed with increased taxes given the budget constrant (2).

The short-run effects on income and the exchange rate of changes in the different policy variables are:

$$dY/dM = (1 + L_2)[T_2 + D(C' + T_1)]/A \lesseqgtr 0$$

$$de/dM = (1 + L_2)(1 - C' - T_1)/A > 0$$

$$dY/dG = D(1 - C' - T_2)(L_1 + L_2/r_o)/A \gtreqless 0$$

$$de/dG = -L_1(1 - C' - T_1)/A < 0$$

It is clear from the above results that when allowance is made for capital gains or losses on the net foreign asset position due to exchange rate changes even the short-run income multipliers for monetary and fiscal policy are of ambiguous sign. The reason for the ambiguity in the case of monetary policy is as follows: the exchange rate depreciation induced by the monetary expansion shifts expenditures towards domestic goods and on this account aggregate demand and income tend to rise; however, if the country is a net debtor in terms of foreign currency, the service of that debt is now higher in terms of domestic goods so disposable income and demand tend to fall on this account. The net effect of the monetary expansion on aggregate demand is therefore ambiguous. For a creditor country disposable income rises with the exchange rate depreciation and this reinforces the expansionary effect on demand.[2]

As previously discussed, any direct effect of fiscal policy on aggregate demand is nullified by an exchange rate appreciation. If, however, the country is a net creditor in the international assets market, the value of its net asset holdings in terms of domestic goods falls which reduces money demand both on the account of lower national income and portfolio size; the fall in money demand reduces the required exchange rate appreciation and thus allows for some domestic expansion. Conversely, for a net debtor country, money demand rises in response to the exchange rate appreciation requiring a further appreciation and fall in aggregate demand.

It follows that fiscal policy will expand or contract income depending on whether the country is a net creditor or debtor in the international assets market. Assuming that initially $D = 0$, which amounts to eliminating the service account and thus making the model identical to the one of Fleming and Mundell, this yields a zero income multipler for fiscal policy which also appreciates the exchange rate. Also, for $D = 0$, the income multiplier for monetary policy is unambiguously positive. Thus, the short-run implications of this model are consistent with those of Fleming and Mundell which were discussed in the previous section.

III. The Long-Run Multipliers

Assuming the existence and stability of a steady state, in the long run the current account will be balanced, thus

(14) $$T(Y + eD - G, e) = -eD$$

Substituting the long-run condition (14) into (8) and (9), the long-run behavior of the model is described by

(15) $$M = L(Y + D', (D' + B)/r_o)$$

[2]In the previous discussion of the adjustment process where no allowance was made for capital gains or losses due to exchange rate changes, the substitution effect of the exchange rate on the trade balance was the only channel through which monetary policy affected aggregate demand. It is, however, possible when $D > 0$ for monetary policy to deteriorate the trade balance while aggregate demand is nonetheless increased. This is so because, other things being equal, the exchange rate depreciation raises disposable income in a creditor country; the rise in Y^D increases consumption demand by more than the deterioration in the trade balance so that aggregate demand rises. The previous argument that a monetary expansion induces a capital outflow and thus tends to improve the long-run service account is still valid if the system is stable. The change in the rate of capital outflow equals that in the current account surplus, or

$$\frac{dT}{dM} + D\frac{de}{dM} = (1 + L_2)(1 - C')(D + T_2)/A$$

which must be positive if the system is stable ($D + T_2 < 0$ implies $A < 0$ so the above would still be positive but the system would be unstable). Similarly, for a debtor country, a fiscal expansion may actually improve the trade balance; given stability, however, the current account must deteriorate.

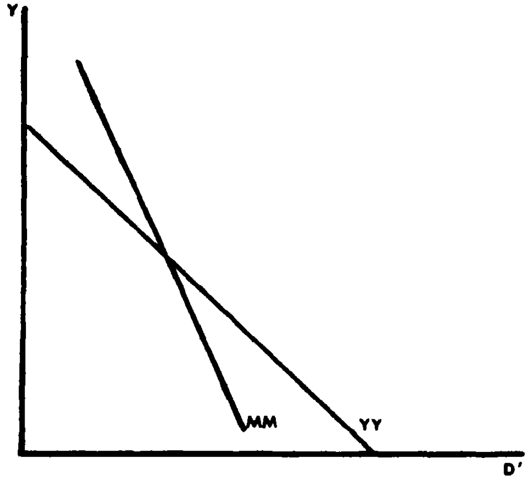


FIGURE 1. LONG-RUN EQUILIBRIUM

and

$$(16) \quad Y = C(Y + D' - G) + G - D'$$

where $D' = eD$, the service account surplus in terms of domestic goods.[3]

Figure 1 illustrates the determination of the long-run equilibrium values of domestic output Y and the service account D'. The downward-sloping schedule YY shows the combinations of Y and D' for which goods market equilibrium prevails. Since from (16) the level of national income, $Y + D'$, is fully determined by the parameters of the consumption function and the level of government spending, the slope of YY is -1 and its intersection with either axis shows the equilibrium level of $Y + D'$. The MM curve corresponds to the locus of monetary equilibrium, its slope is steeper than -1 since a higher D' raises money demand on account of the portfolio effect requiring a more than proportional fall in Y in order to reduce national income, $Y + D'$, and thus reduce money demand for restoring equilibrium.

A fiscal expansion ($\Delta G > 0$) raises national income in proportion to the standard Keynesian multiplier, the YY locus is thus shifted rightwards while MM remains unchanged. In the new equilibrium, Y is necessarily higher while the service account deteriorates. A monetary expansion shifts MM rightwards such that, with an unchanged national income, domestic output falls by the full amount of the improvement in the service account. The long-run results, therefore, verify the conjecture expressed in the first section that the long-run output multiplier for monetary policy is negative while it is positive for fiscal policy. These results, of course, stand in sharp contrast with the short-run results where, for $D = 0$, fiscal policy was unable to affect either domestic or national income while expansionary monetary policy would increase both. The above results can be algebraically verified by differentiating (15)

[3]Notice that if $L(\cdot)$ does not depend on the stocks of other assets held, each of (15) and (16) would require a given level of national income to equilibrate the respective markets; unless those two required levels were by chance equal, a long-run solution would not exist. Also, the dependence of money demand on income is crucial for the stability of the long-run solution. If $L_1 = 0$, it follows from (13) that the system would be stable or unstable depending solely on whether the country is a net creditor or debtor; $L_1 > 0$ allows for the possibility of stability even for a debtor country. Notice also that if the traded bond were to be denominated in terms of domestic currency there would be no capital gains or losses due to exchange rate changes and therefore the term $D[L_1 + (1 - C' - T_1)L_2/r_o]$ in (13) would not appear; in this case the stability condition becomes $d\dot{D}/dD = -L_1(1 - C')/L_2 < 0$ which is satisfied given our prior assumptions about signs

and (16):

$$dY/dM = -(1 + L_2)r_o/L_2 < 0$$
$$dD'/dM = (1 + L_2)r_o/L_2 > 0$$
$$dY/dG = 1 + r_oL_1/L_2 > 1$$
$$dD'/dG = -r_oL_1/L_2 < 0$$

and for national income:

$$d(Y + D')/dM = 0$$
$$d(Y + D')/dG = 1$$

In this model the long-run equilibrium level of national income is fully determined by the standard Keynesian fiscal policy multiplier and the interest rate, as in a closed economy. In contrast to a closed economy, however, the existence of perfect capital mobility implies that monetary policy has no leverage on the interest rate and is thus unable to affect national income. Any effect of monetary policy on domestic income (and employment), therefore, depends solely on its effect on the service account (since $Y + D'$ is independent of M). A monetary expansion induces a portfolio shift towards foreign assets and the service account is improved; domestic income therefore must fall by the full amount of the improvement in the service account. A balanced expansion in government spending and taxes raises national income in the same proportion as the budget increases; since in addition the service account deteriorates, domestic income must increase more than proportionately.

REFERENCES

J. M. Fleming, "Domestic Financial Policies under Fixed and under Flexible Exchange Rates," *Int. Monet. Fund Staff Pap.*, Nov. 1962, *9*, 369–79.

Duncan Foley and Miguel Sidrauski, *Monetary and Fiscal Policy in a Growing Economy*, New York 1971.

R. Mundell, "Capital Mobility and Stabilization Policies under Fixed and Flexible Exchange Rates," *Can. J. Econ.*, Nov. 1963, *29*, 475–85.

Don Patinkin, *Money, Interest, and Prices*, New York 1965.

Hedonic Theory and the Demand for Cable Television

By Bryan Ellickson*

The techniques of hedonic price theory are used in this paper to derive demand equations for cable television.* This application of hedonic theory requires a modification of the standard model (as formulated by Sherwin Rosen) because consumers are not choosing among a continuous spectrum of commodities. Cable television can be regarded as an indivisible commodity, but the only option open to the consumer is either to subscribe to the cable system serving his community or not. Thus the most familiar feature of hedonic theory, the tangency of hedonic and bid-price curves, has no role to play in the present context. But other features of the hedonic approach, the description of indivisible commodities in terms of characteristics and representation of consumer preference by bid-price functions, do prove useful in modeling the demand for cable. The approach developed here should have an immediate application in a number of other contexts involving estimation of demand equations for an indivisible commodity where the only choice open to the consumer is to subscribe to the service or not.

The basic model, presented in Section I, can be regarded as a generalization of the theory of cable demand set forth by Roger Noll, Merton Peck, and John McGowan (hereafter, N-P-M). As an illustration of this generality, the theory is used to derive the cable penetration equation estimated by Rolla Park. The theoretical framework permits a very natural comparison of the economic content underlying the specifications chosen by N-P-M and Park, in sharp contrast to the difficulty of comparing their equations in their original form.

The demand equations derived in Section I are not, strictly speaking, appropriate to the problem addressed by N-P-M and Park since all consumers in a cell are assumed to have the same income while N-P-M and Park estimate these equations using television markets as the unit of observation. Section II demonstrates how the theory can be used to derive market demand equations when incomes are lognormally distributed. In practice estimation of this version of the model will produce essentially the same results as the logistic employed by Park.

One of the most intriguing uses N-P-M find for their theory of cable demand is the estimation of the value consumers attach to additional television services. In Section III the same approach is applied to Park's model in order to shed some light on how seriously N-P-M's estimates of the value of a fourth network should be taken. While Park's model provides a better fit to the data than does that of N-P-M, his estimates imply a value for a fourth network that is unreasonably high. Furthermore, relatively small variations in the estimated parameters lead to dramatic changes in this estimate. The conclusion drawn here is that this use of the theory simply asks too much of the data.

I. Modeling the Demand for Cable

Consider a cable system serving an area populated by N households. Assume that household utility can be represented by the functions $U_i(x_i, q_j)$ where x_i is an n-dimensional vector of private goods consumed by the ith household and q_j is an m-dimensional vector of characteristics associated with the jth television mode. In practice the components of q_j include the number of

*Associate professor, University of California-Los Angeles. This paper was written under a grant from the John and Mary R. Markle Foundation to the Communications Policy Program at the Rand Corporation. I am grateful to Julie Da Vanzo, Roger Noll, and Rolla Park for comments on an earlier draft.

television signals of different types (network, independent, educational) and measures of signal quality. I consider here the case where only two modes are available: the household can choose either to watch television over the air with characteristics q_o or over cable with characteristics q_c. From the point of view of the consumer the characteristics associated with each mode are exogenous. Each utility function is assumed to be a strictly quasi-concave function of the x_i for given q_j.

The key to my approach is the formulation of consumer choice as a two-stage process. Let us assume that a household selecting the jth television mode will choose x_i to maximize $U_i(x_i, q_j)$ subject to the budget constraint $px_i + S_j = y_i$ where $p = (p_1 \ldots p_n)$ is a vector of prices, y_i is the household's money income, and S_j is the cost to the household of watching the jth television mode. The solution to this maximizing problem can be characterized in terms of an indirect utility function

$$V_i = \phi_i(p, q_j, y_i - S_j)$$

where V_i represents the maximum utility attainable by the ith household if it watches the jth television mode. Also assume nonsatiation so that V_i is a strictly increasing function of $y_i - S_j$.

In choosing whether to watch television over the air or over cable, the household will select the mode j for which utility $\phi_i(p, q_j, y_i - S_j)$ is maximized. Ignoring the option to own no television set at all, we can without loss of generality set $S_o = 0$ and $S_c = S$ where S is the subscription fee for cable. Then the ith household will be indifferent between watching television over the air or over cable provided that

$$\phi_i(p, q_o, y_i) = \phi_i(p, q_c, y_i - S_i^*) \tag{1}$$

where S_i^* is the maximum fee that the household will pay for cable. Solving (1) explicitly for S_i^*,

$$S_i^* = \psi_i(p, q_o, q_c, y_i) \tag{2}$$

the *bid-price function* for the ith consumer. Assuming that when provided free of charge cable service is preferred to reception over the air, $0 \le S_i^* \le y_i$. It is also easy to show that ψ must be homogeneous of degree one in p and y_i.

Equation (2) is the basic relationship that will be used to derive the demand for cable. Suppose that the bid prices S^* of consumers in a market are distributed according to some probability law with density $h(S^*)$. If S is the fee actually charged for the cable service, a household will subscribe[1] if and only if $S^* \ge S$. Letting N_c denote the number of cable subscribers, cable penetration $n = N_c/N$ is given by

$$n = \int_S^\infty h(S^*)dS^* \tag{3}$$

The model developed by N-P-M can be used to illustrate the approach outlined above. Assuming that x_i and q_j are scalars, they adopt a Cobb-Douglas form for the utility function, $U_i = x_i q_j^\gamma$, where γ is a constant. It is easy to show that bid price is given by

$$S_i^* = [1 - (q_o/q_c)^\gamma]\, y_i \tag{4}$$

corresponding to equation (A-2) of N-P-M. Noll, Peck, and McGowan assume that all consumers have the same income, $\bar{y}$. The taste parameter γ is assumed to be randomly distributed with probability density $\mu e^{-\mu\gamma}$ where μ is a positive constant. Using the standard technique for change of variable, we can translate this assumption into the following expression for the probability density of bid price:

$$h(S^*) = \frac{1}{\theta(\bar{y} - S^*)} \cdot \exp\left[\frac{1}{\theta} \log(1 - S^*/\bar{y})\right] \tag{5}$$

where $\theta = (1/\mu)\log(q_c/q_o)$. Performing the integration indicated by equation (3), we obtain

$$\frac{\log(1 - S/\bar{y})}{\log(n)} = \frac{1}{\mu}\log(q_c/q_o) \tag{6}$$

[1]Households for which $S^* = S$ will be indifferent between television over the air and on cable, but with the continuous probability densities we employ the set of such households will have measure zero.

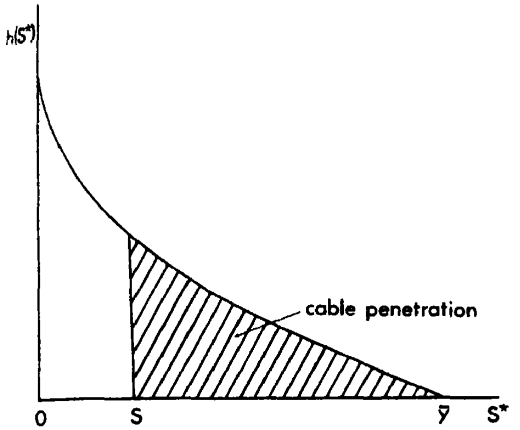


FIGURE 1

which corresponds to the cable penetration equation (A-8) estimated by N-P-M. This translation of the assumptions employed by N-P-M into a statement about the distribution of bid prices permits their model to be given a natural economic and graphical interpretation. If $\theta < 1$ (which is the case in N-P-M's empirical estimates), then the density $h(S^*)$ decreases monotonically with S^* in the interval $[0, \bar{y}]$ and equals 0 elsewhere. This is illustrated in Figure 1 where the shaded area represents the penetration achieved for a given subscription fee S.

To illustrate how the theory presented above can be used to generate alternatives to the N-P-M model, I will now show how, under the assumption that all households in a given market have identical income $\bar{y}$, the equation employed by Park can be derived. Assume that the distribution of bid prices in the market can be described by the equation

$$(7) \qquad \log S^* = \psi(p, q_o, q_c, \bar{y}) + \epsilon$$

where ϵ has a logistic distribution with mean 0 and variance σ^2. Then $\log S^*$ has a logistic distribution with mean $\mu = \psi(p, q_o, q_c, \bar{y})$ and variance σ^2. The probability density can be written as

$$(8) \qquad h(z^*) = \frac{\exp(z^*)}{\tau[1 + \exp(z^*)]^2}$$

where $z^* = (\log(S^*) - \mu)/\tau$ and $\tau = \sqrt{3}\sigma/\pi$ (see David Cox, p. 101). If a subscription fee S is charged for cable, then the penetration rate is

$$(9) \qquad n = prob\{S^* \geq S\} = \int_z^\infty h(z^*)dz^* = \frac{1}{1 + \exp(z)}$$

where $z = (\log(S) - \mu)/\tau$. By appropriate manipulation, equation (9) can be written as

$$(10) \qquad \log\left(\frac{n}{1-n}\right) = \frac{\mu}{\tau} - \frac{1}{\tau}\log S = \frac{1}{\tau}\psi(p, q_o, q_c, \bar{y}) - \frac{1}{\tau}\log S$$

Park's penetration equation can then be given a direct interpretation in terms of the implied distribution of bid prices. If we assume that

$$(11) \qquad \log S^* = f(q_o, q_c) + \alpha \log \bar{y} + \epsilon$$

where $f(q_o, q_c)$ is a function of the vectors q_o and q_c that is linear in the parameters, then equation (10) takes the form

$$(12) \qquad \log\left(\frac{n}{1-n}\right) = \frac{1}{\tau} f(q_o, q_c) + \frac{\alpha}{\tau}\log \bar{y} - \frac{1}{\tau}\log S$$

which is precisely the equation estimated by Park.

As illustrated in Figure 2, the probability density function $h(S^*)$ in Park's model is a unimodal skewed distribution.[2] The shaded area represents the penetration achieved for a given subscription fee S.

A direct comparison of the models estimated by N-P-M and Park is not an easy task: equations (6) and (12) appear to have little in common. However, by translating the models into their implied distributions for S^*, exploring their relationship becomes a straightforward matter. The density $h(S^*)$ in Park's model differs most sharply from

[2]The density $h(S^*)$ is derived from the density $h(\log S^*)$ by the standard technique for change of variable: i.e., $h(S^*) = h(\log S^*)/S^*$. I have used the same function sign for both densities to avoid complicating the notation.

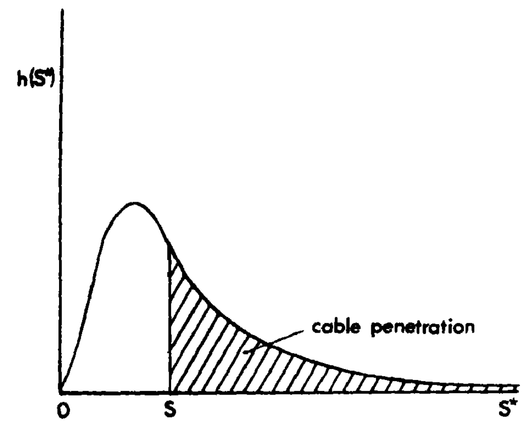


FIGURE 2

that of N-P-M in the region to the left of the mode. Setting the first derivative of $h(S^*)$ equal to zero, the mode $\hat{S}^*$ is determined by the expression $log\ \hat{S}^* = \mu + \tau\ log(1 - \tau/1 + \tau)$ which, when substituted into equation (10), yields $log(\hat{n}/1 - \hat{n}) = -log(1 - \tau/1 + \tau)$ where $\hat{n}$ is the penetration when $S = \hat{S}^*$. In what Park refers to as his preferred equation (p. 143, equation (*)) the coefficient of $log\ S$ is $1/\tau = 1.473$, implying a penetration $\hat{n} = .84$ at the mode, much higher than the penetration rates observed in the sample (which averaged about .3; see Park, p. 140, fn. 19). Thus, the density functions implied by the empirical results of Park and N-P-M resemble one another at least to the extent that both decline monotonically with S^* over the empirically relevant range.

So far I have said nothing about how equation (10) should be estimated. Fortunately, a method developed by Joseph Berkson for the binary logit model is directly applicable to the present case. If the variable n in equation (10) is interpreted as the probability that a household in the given market will subscribe to cable, then N_c, the number of households in the market who actually subscribe to cable, will be binomially distributed with $E(\tilde{n}) = E(N_c/N) = n$ where $\tilde{n}$ is the *observed* penetration rate. Writing equation (9) as $n = \tilde{H}(z)$ where $\tilde{H}(z) = 1 - H(z)$ and $H(z)$ is the logistic cumulative density function (*cdf*), a Taylor series expansion of the inverse transformation $z = \tilde{H}^{-1}(n)$ about the observed penetration rate n yields the equation

$$(13) \quad log\left(\frac{\tilde{n}}{1-\tilde{n}}\right) = \frac{1}{\tau}\psi(p_o, q_o, q_c, \bar{y}) - \frac{1}{\tau}\ log\ S + \epsilon$$

where we have made the substitutions $\tilde{H}^{-1}(\tilde{n}) = -log(\tilde{n}/1 - \tilde{n})$ and

$$\tilde{H}^{-1}(n) = \frac{1}{\tau}\ log\ S - \frac{1}{\tau}\psi(p, q_o, q_c, \bar{y})$$

and denoted the remainder term by $-\epsilon$. The disturbance ϵ is binomially distributed with zero mean; if the number of households in each television market is large, it will be approximately normally distributed. Therefore, ordinary least squares applied to equation (13) will yield consistent estimates of the parameters.[3]

II. A Market Demand Equation for Cable

It is worth noting at this juncture that there is an alternative way to derive Park's equation from a theory of utility maximization, the analysis of quantal choice developed by McFadden (1974). The contents of this paper thus far can be viewed, therefore, as an alternative motivation for the binary logit model. However, there is an aspect of the empirical problem faced by N-P-M and Park which has not been addressed in the literature on quantal response. Just as with the analysis of McFadden, the models discussed in the preceding section are more properly viewed as theories of individual rather than market response where $\bar{y}$ represents an individual household's income and n is interpreted as the probability that a particular household will subscribe to

[3]The preceding argument is a direct adaptation of David McFadden (1976). As he notes, the disturbance term will typically be heteroskedastic, and weighted least squares are required in order to get asymptotically efficient estimates.

cable. But it is frequently the case in problems of this sort that market level data are all that is available. Furthermore, even when data can be obtained for individual households, it is often the market response which is more relevant to policy-oriented research. Applying these models to market data requires grouping households together who reside in a given market, a procedure that will result in aggregation bias to the extent that incomes within the market are not identical. Therefore, in this section I present an alternative model yielding a market demand for cable equation in which the restriction that all households have the same income is relaxed. While the form of the resulting equation is not logistic, it turns out to be approximately equivalent to the equation estimated by Park under conditions likely to be encountered in practice.

Turning again to the bid price equation (11) implicit in Park's model, we have

$$(14)\quad \log S^* = f(q_o, q_c) + \alpha \log y + \epsilon$$

where $\bar{y}$ is replaced by y to indicate that household incomes are no longer assumed identical. Assume that within a given market $\log y$ and ϵ are independent and normally distributed, $N(\mu_1, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively. Then $\log S^*$ will be normally distributed with mean $\mu = f(q_o, q_c) + \alpha\mu_1$ and variance $\sigma^2 = \alpha^2\sigma_1^2 + \sigma_2^2$, and $z^* = (\log S^* - \mu)/\sigma$ will be distributed $N(0, 1)$. If the cable subscription fee is S, then cable penetration will be given by

$$(15)\quad n = \frac{1}{\sqrt{2\pi}} \cdot \int_z^\infty \exp[-(z^*)^2/2]dz^* \equiv \tilde{H}(z)$$

with $z \equiv (\log S - \mu)/\sigma$ and $\tilde{H}(z) \equiv 1 - H(z)$ where $H(z)$ is the standard normal *cdf*. Taking the inverse, the transformation used in probit analysis,

$$(16)\quad -\tilde{H}^{-1}(n) = -z = \frac{\mu}{\sigma} - \frac{1}{\sigma}\log S$$

where I have multiplied through by -1 in order to facilitate comparison of this equation with that of Park. Substituting in the expression for μ, and writing μ_1 as $\overline{\log y}$, we obtain

$$(17)\quad -\tilde{H}^{-1}(n) = \frac{1}{\sigma} f(q_o, q_c) + \frac{\alpha}{\sigma}\overline{\log y} - \frac{1}{\sigma}\log S$$

an equation that looks very similar to equation (12) derived for Park's model in the preceding section. The argument presented at the end of Section I with minor modification provides a justification for estimating this equation by ordinary least squares.

The density function $h(S^*)$ in this model is lognormal, with a skewed shape resembling that portrayed in Figure 2. The similarity between this model and that of Park is, in fact, much closer than the analysis so far would suggest. It is a well-established fact that, except in the extreme tails, the numerical difference between the normal and the logistic *cdf* is very small (see Winifred Ashton, p. 11, or Cox, p. 28). Since the right-hand sides of equations (12) and (17) are essentially the same apart from multiplication by a scalar,[4] this claim is equivalent to the assertion that the dependent variables will be highly correlated. Taking $n = .3$, approximately the average penetration rate observed in Park's sample, we find that $\log(n/1 - n) = -.8473$ and $-\tilde{H}^{-1}(n) = -.5244$ so that the dependent variable in the logistic model differs from that of the lognormal model by a factor of 1.616. Using this factor to adjust the dependent variables to a common scale, the close

[4]There is one discrepancy: $\overline{\log y}$ in equation (17) replaces $\log \bar{y}$ of equation (12) where $\bar{y}$, in the empirical study of Park, is the average household income in the television market. Using a well-known relationship for the lognormal distribution (see John Aitchison and James Brown, p. 8), $\bar{y} = \exp(\overline{\log y} + \sigma_1^2/2)$ where σ_1^2 is the variance of $\log y$. Thus $\overline{\log y} = \log \bar{y} - \sigma_1^2/2$. If σ_1^2 is invariant across television markets, then $\log \bar{y}$ can be used in place of $\overline{\log y}$ with the term $-\alpha\sigma_1^2/2$ absorbed into the constant term; if not, a specification error may occur which does not seem unlikely since σ_1^2 is probably correlated with $\log \bar{y}$.

TABLE 1

n	$log(n/1-n)$	$-1.616\tilde{H}^{-1}(n)$
.65	.6190	.6226
.60	.4054	.4095
.55	.2007	.2031
.50	0	0
.45	−.2007	−.2031
.40	−.4054	−.4095
.35	−.6190	−.6226
.30	−.8473	−.8474
.25	−1.0986	−1.0900
.20	−1.3863	−1.3600
.15	−1 7346	−1.6748
.10	−2.1972	−2.0711

relationship between the two is illustrated in Table 1.

III. The Value of Television

One of the most intriguing uses which N-P-M find for their model is the estimation of the value consumers attach to television services, and in particular, the value which would be derived from the addition of a fourth network. The procedure employed by N-P-M has a simple interpretation within the theoretical framework developed in Section I. The average value of cable service is just the mean of the bid-price distribution $E(S^*) = \bar{y}\theta/(1+\theta)$ where $\theta = (1/\mu)\ log\ (q_c/q_o)$. When multiplied by the number of households, N, this gives an estimate of total "consumer's surplus," $W = N\bar{y}\theta/(1+\theta)$, corresponding to N-P-M's equation (A-15). By assuming that nothing is available over the air, N-P-M are able to use their estimate of equation (6) to determine the value of θ for various combinations of television signals carried over cable. The first column of Table 2 lists the estimates of consumer surplus reported by N-P-M for varying numbers of primary network stations as a percentage of average income $\bar{y}$ (see N-P-M, p. 288, Table A-2).

TABLE 2

Number of Primary Network Stations	Estimated Consumer Surplus (percent of mean income)	
	N-P-M	Park
1	2.60	2.08
2	4.06	4.89
3	5.07	8.95
4	5.83	14.30
5	6.45	20.99

An analogous procedure can be applied to Park's model. Using the probability density function (8) and assuming[5] that $\tau < 1$,

$$(18)\quad E(S^*) = \int_{-\infty}^{\infty} \frac{S^* \exp(z^*)}{\tau[1+\exp(z^*)]^2}\, dz^* = \exp(\mu)B(1+\tau, 1-\tau) = \frac{\exp(\mu)\tau\pi}{sin(\tau\pi)}$$

where $B(\ )$ is a *beta* function. In Park's preferred equation $1/\tau = 1.473$ so that $\tau\pi/sin(\tau\pi) = 2.520$. Adopting N-P-M's assumption that no signals are available over the air, setting average household income equal to $10,000 and letting the fraction of households owning color sets equal .5 (approximately the sample means), Park's preferred equation can be used to estimate μ in equation (18) for varying numbers of primary network stations. The estimated consumer's surplus as a fraction of mean household income is given in the second column of Table 2.

The resulting estimates are not too dissimilar from those of N-P-M for one or two network signals, but as the number of signals is increased the estimated consumer surplus implied by Park's equation becomes increasingly unrealistic (with consumers willing to pay almost 21 percent of their income to receive five primary network signals). Calculation of the value of a fourth network leads to similarly doubtful estimates in Park's model: relative to a situation with three primary network signals and one independent, N-P-M estimate that consumers would be willing to pay an addi-

[5]The condition that $\tau < 1$ is equivalent to requiring the coefficient of $log\ S$ in equation (12) be greater than one (in absolute value) which it is in the estimates reported by Park.

tional .76 percent of their income to receive one more network signal (see N-P-M, p. 30) while Park's equation implies that the additional amount would be 5.92 percent. Thus, compared to the N-P-M estimate of $4.2 billion, Park's model suggests a total of $32.7 billion as the value of a fourth network.

The estimates of consumer's surplus implied by Park's model are clearly quite unreasonable. But the major lesson to be learned from this exercise is that estimates of this sort, whether based on the model of Park or of N-P-M, are not to be trusted. In the first place, the calculations involve extrapolation of the empirical results far beyond conditions warranted by the sample data.[6] Three primary networks and an independent available over cable with no signals received over the air would imply a penetration rate $n = .94$ for Park's preferred equation assuming a subscription fee of $63 (the sample mean), while the average penetration in his sample was only .3 (in N-P-M's model, the penetration rate would be .90). Furthermore, the calculation of consumer surplus for Park's model is extremely sensitive to the estimate of the parameter τ: if the coefficient of $\log S$ in equation (12) is shifted up or down by half a standard deviation, the percent of mean income consumers are willing to pay for three primary network signals and an independent varies from 2.5 percent (half of what N-P-M estimate) to more than 100 percent. Finally, the estimates of consumer's surplus appear to be very sensitive to the form of the probability density function assumed for S^*. This can be seen perhaps most clearly when Park's results are interpreted in terms of the lognormal model developed in Section II: calculation of $E(S^*)$ for that model leads to estimates of consumer surplus approximately 28 percent lower than those implied by the logistic model even though the estimated coefficients are assumed to be the same.

[6]In commenting on their results, N-P-M reach essentially the same conclusion (see pp. 22–23)

REFERENCES

John Aitchison, and James Brown, *The Lognormal Distribution*, Cambridge 1957.

Winifred Ashton, *The Logit Transformation*, New York 1972.

J. Berkson, "Application of the Logistic Function to Bio-Assay," *J. Amer. Statist. Assn.*, Sept. 1944, *39*, 357–65.

David Cox, *The Analysis of Binary Data*, London 1970.

D. McFadden, "Conditional Logit Analysis of Qualitative Choice Behavior," in Paul Zarembka, ed., *Frontiers in Econometrics*, New York 1974, 105–43.

———, "Quantal Choice Analysis: A Survey," *Annals Econ. Soc. Measure.*, Fall 1976, *5*, 363–90.

Roger Noll, Merton Peck, and John McGowan, *Economic Aspects of Television Regulation*, Washington 1973.

R. Park, "Prospects for Cable in the 100 Largest Television Markets," *Bell J. Econ.*, Spring 1972, *3*, 130–50.

S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 34–55.

On Regulation and Uncertainty: Comment

By NICHOLAS RAU*

In a recent article in this *Review*, Yoram Peles and Jerome Stein make the interesting point that "The Effect of Rate of Return Regulation is Highly Sensitive to the Nature of the Uncertainty." By this they mean that the well-known proposition of Harvey Averch and Leland Johnson (A-J) about the effect of rate of return regulation on a monopolist's scale of plant may or may not hold in an uncertain environment, depend-on how the state of nature affects the maximal quasi-rent function.

The first three sections of this paper present a general, self-contained analysis of the problem, whose purpose is to elucidate the economics of Peles and Stein's findings. My main conclusion may be summarized as follows: if the state of nature affects both the average and the marginal return on capital, then whether an A-J or anti A-J effect prevails depends on the amount of randomness in the environment. The more "noise," the more likely is an anti A-J effect.

Actually there are two issues here, which I shall refer to as the "global problem" and the "local problem." The global problem concerns a comparison of the behavior of the regulated monopolist with that of his unregulated counterpart. The mathematics of this problem are set out in Section I and the economic implications are discussed in Section II. The local problem concerns the way in which a regulated monopolist reacts to a small change in the maximal allowable rate of return. This problem is discussed rather briefly in Section III. I derive a formal solution which is somewhat less than iridescent in its economic interpretability, but serves to generate a simple example illustrating the main points of my paper.

*Lecturer in political economy, University College London, England. I would like to thank A. B. Atkinson and G. F. Mathewson for helpful comments on an earlier draft.

In Section IV, I return to the original analysis of Peles and Stein, and take issue with one of their main results. This is the "anti A-J theorem for multiplicative uncertainty." What seems to be the most precise statement of this "theorem" is given in their article:

> If, however, the uncertainty is multiplicative in the maximal quasi-rent function, then a decline in the regulated rate of return towards the cost of capital will lower the optimum *ex ante* scale of plant relative to what would be chosen if there were no regulation. [p. 284]

This statement can be given two alternative interpretations, neither of which is true in general. I shall give a (fairly weak) sufficient condition for one of these interpretations to be correct.

I. The Global Problem: Analysis

Consider a regulated monopolist whose maximal quasi-rent function, gross of the "tax" effectively imposed by the regulatory authorities, is $R(K,u)$. Here K denotes capital stock and u is a random variable representing the "state of nature." We assume

$$R_2(K,u) > 0 \tag{1}$$

$$R_{11}(K,u) < 0 \tag{2}$$

Throughout this paper, numerical subscripts denote partial differentiation (double subscripts denote second partials). The inequality (1) states that states of nature can be ranked in such a way that the larger value of u implies a higher quasi rent R at any given value of capital stock. The inequality (2) states that R is a strictly concave function of K for given u.

Postulate that u is distributed over some interval $[a,b]$ of the real line with absolutely

continuous cumulative distribution function F and density function f.

A word of warning should be inserted at this stage. Suppose that $R(0,a) < 0$: this raises the possibility of a shut-down option (with associated nonconvexities) and also the possibility that for some (K,u) pairs, the average revenue product of capital may be positive but *less than* the marginal revenue product of capital. In what follows we shall simply ignore both these possibilities. We therefore make the simplifying assumption that all (K,u) pairs which play a crucial role in the sequel have the property that

$$R(K,u) > K \max [0, R_1(K,u)] \tag{3}$$

Let s be the maximal rate of return allowed by the regulatory authorities. Applying the inequality (1), we may associate with each pair (K,s) a unique *critical state of nature* $v(K,s)$ defined as follows:

$$v(K,s) = a \text{ if } R(K,a) > sK \tag{4a}$$

$$v(K,s) = b \text{ if } R(K,b) < sK \tag{4b}$$

$$R(K, v(K,s)) = sK \text{ if } R(K,a) \leqq sK \leqq R(K,b) \tag{4c}$$

Equation (4c) is interpreted as follows. Suppose that, given K and s, there is a value of the random variable u which makes the gross rate of return $R(K,u)/K$ exactly equal to s. Then this value of u is called the critical state of nature and is denoted by $v(K,s)$. The relations (4a) and (4b) extend this definition of $v(K,s)$ in an obvious fashion to the cases where the above supposition is false.

Let the cost of capital be i, where $0 < i < s$. Then, assuming risk neutrality, the monopolistic chooses K so as to maximize the expression

$$\int_a^{v(K,s)} R(K,u)\, f(u)du + sK(1 - F(v(K,s))) - iK$$

Let the optimal value of K be K_s and let $v(K_s,s) = v_s$. Assume that the optimum is an interior one in both relevant senses: $K_s > 0$ and $a < v_s < b$.

Then

$$\int_a^{v_s} R_1(K_s,u)\, f(u)du = i - s(1 - F(v_s)) \tag{5}$$

$$R(K_s, v_s) = sK_s \tag{6}$$

Now consider the *unregulated* monopolist in otherwise identical circumstances. This entrepreneur wishes to maximize the expression

$$\int_a^b R(K,u)\, f(u)du - iK$$

Letting the unregulated optimal capital stock be K_∞, and again assuming an interior optimum, we have

$$\int_a^b R_1(K_\infty,u)\, f(u)du = i \tag{7}$$

From (5) and (7) we have

$$\int_a^b [R_1(K_\infty,u) - R_1(K_s,u)]\, f(u)du = s(1 - F(v_s)) - \int_{v_s}^b R_1(K_s,u)\, f(u)du \tag{8}$$

In view of the strict concavity assumption (2), the left-hand side of (8) has the same sign as $(K_s - K_\infty)$. Also, by elementary probability theory,

$$1 - F(v_s) = \int_{v_s}^b f(u)du$$

Thus (8) implies that

$$K_s - K_\infty \sim \int_{v_s}^b [s - R_1(K_s,u)]\, f(u)du \tag{9}$$

Where the symbol $\sim$ means "has the same sign as." The relations (6) and (9) are the crucial results for our investigation of "global" A-J and anti A-J possibilities.

II. The Global Problem: Implications and Interpretations

An immediate consequence of (9) is the following:

PROPOSITION: *A sufficient, but not necessary, condition for K_s to be greater than K_∞ is that*

$$s > R_1(K_s,u) \qquad \textit{for all } u > v_s$$

To interpret this proposition, we make use of the simplifying assumption intro-

duced in the previous section. Assuming, then, that (3) holds at $(K, u) = (K_s, v_s)$ we have $R(K_s, v_s)/K_s > R_1(K_s, v_s)$.

Applying (6), we have $s > R_1(K_s, v_s)$. This gives us an immediate corollary to the proposition.

COROLLARY: *If $R_1(K, u)$ does not depend on u, then $K_s > K_\infty$.*

This corollary is the "A-J theorem for additive uncertainty" of Peles and Stein: if $R(K, u)$ is of the form "function of K *plus* function of u," then $K_s > K_\infty$.

From now on, let us assume that

$$R_{12}(K, u) > 0 \qquad (10)$$

which states that, given any capital stock K, the ranking of states of nature with respect to total quasi rent is identical with the ranking with respect to marginal quasi rent. Peles and Stein's "multiplicative uncertainty" (where $R(K, u)$ is the product of a function of K alone and a function of u alone) is a special case of this.

Given (10), we may interpret the proposition as follows: if there is *no* state of nature so bountiful that the *marginal* quasi rent R_1 at the regulated optimal capital stock and that state of nature exceeds s, then the A-J inequality ($K_s > K_\infty$) must obtain; if on the other hand $R_1(K_s, b) > s$, then K_s may be greater or less than K_∞.

This analysis suggests that Peles and Stein's emphasis on the multiplicative case is slightly misleading. For suppose that we specify a particular functional form for $R(K, u)$ satisfying (10): then this information alone is simply insufficient to sign $(K_s - K_\infty)$. To decide whether A-J or anti A-J is the case, we also need information about the distribution of u.

To highlight this point, let us first consider a general thought experiment, and then work through a particular example. Suppose we have a given quasi-rent function R and a distribution function F for u. Given also i and s, the regulated monopolist's capital stock K_s and critical state of nature v_s are determined as above, as is also the unregulated capital stock K_∞. Now suppose that keeping everything else unaltered, we change the distribution function F by shifting probability mass *within the range* $u > v_s$. It is immediate from (5) that this will have no effect on the regulated monopolist's behavior. The unregulated monopolist, however, will choose (in general) a different K_∞. Indeed suppose that (10) holds and that there exists a state w such that $v_s < w < b$ and $R_1(K_s, w) = s$. If most of the probability mass of u to the right of v_s is concentrated in the range $v_s < u < w$, then the A-J result will obtain: $K_s > K_\infty$. If, on the other hand, most of the probability mass to the right of v_s is concentrated in the range $w < u < b$, then we will have the anti A-J result $K_s < K_\infty$.

The particular example which I now discuss is not quite in the spirit of the above analysis, since I consider a *discrete* distribution of states of nature, but the general principle still holds. Consider the case of regulation under *certainty* illustrated in Figure 1. There the quasi-rent function is $R(K)$, K_s is the regulated optimum, and K_∞ is the unregulated optimum. Evidently, $R(K_s)/K_s = s > i$ and $R'(K_s) < s$. These regulations are not inconsistent with the inequalities

$$0 < R'(K_s) < 2i - s \qquad (11)$$

For reasons to be made plain shortly we assume that (11) holds.

Now introduce uncertainty as follows. Suppose that there is now a probability 1/2 of the quasi-rent function being R, and a probability 1/2 of the quasi-rent function

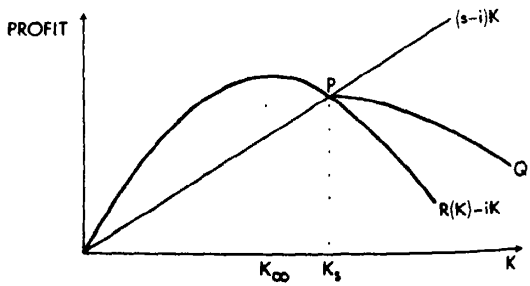


FIGURE 1

being T, where $T(K) > R(K)$ and $T'(K) > R'(K)$ for all K.

Given (11), the new expected net profit function for the regulated monopolist *must* look like the curve *OPQ* in Figure 1. Thus the change in environment is no help to the regulated monopolist. On the other hand the *unregulated* monopolist will move to a capital stock K_n which exceeds K_∞ and *may* exceed K_s.

Suppose, in fact, that $T(K) = (1 + \epsilon)\cdot R(K)$, where $\epsilon > 0$. Then K_n is given by the equation $(2 + \epsilon)R'(K_n) = 2i$. Thus $K_s > K_n$ if and only if $2i/(2 + \epsilon) > R'(K_s)$. By (11), $R'(K_s) < 2i - s < i$. Thus if ϵ is small, $K_s > K_n$. On the other hand, if ϵ is very large then $K_s < K_n$. Thus a small value of ϵ implies an A-J effect, and a large value of ϵ an anti A-J effect. This is fully consonant with my remark above about "bountiful" states of nature.

III. The Local Problem

Let K_s be as in Section I. We are interested in how K_s reacts to small changes in s. To perform this analysis, it is helpful to adopt a rather different mathematical technique from that used in Section I. Instead of working with the distribution function of the random variable u, let us work directly with the distribution of the *ex post* rate of return.

For any capital stock K and any rate of return r, we set

$$G(K,r) = Prob(R(K,u) > rK)$$

Then the regulated firm's maximand may be written[1]

$$H(K,s) = \left[\int_0^s G(K,r)dr - i\right]K$$

Assuming as before that we have an interior maximum, we have $H_1(K_s,s) = 0$. Differentiation with respect to s yields

$$H_{11}(K_s,s)\cdot(dK_s/ds) + H_{12}(K_s,s) = 0$$

Assuming a regular interior maximum, we have $H_{11}(K_s,s) < 0$. Now

$$H_{12}(K,s) = G(K,s) + KG_1(K,s)$$

for all s. Hence

$$(12)\quad dK_s/ds \sim G(K_s,s) + K_sG_1(K_s,s)$$

The statement (12) is the "formal solution" mentioned in my introductory comments. As noted there, (12) is not easy to interpret in terms of economics; notice in particular the nasty possibility that K_s may not be monotonic in s. However, one rather instructive special case is easily analyzed via (12).

Let α and λ be positive numbers, with $\lambda > 1 > \alpha$. Now suppose that $R(K,u) = K^\alpha(1 + u)$, where $F(u) = 1 - (1 + u)^{-\lambda}$ for $u \geqq 0$.

Then

$$G(K,r) = \min[1,(rK^{1-\alpha})^{-\lambda}]$$

Given $s > i$, K_s will be chosen such that[2]

$$sK_s^{1-\alpha} > 1$$

Thus, from (12),

$$dK_s/ds \sim 1 - \lambda(1 - \alpha)$$

This is fully in accord with the global results of Section I. A large value of λ implies that $dK_s/ds < 0$ (the A-J case): and a large value of λ implies a rather unnoisy environment. On the other hand, if λ is small, then $f(u)$ has a fat upper tail, and $\lambda(1 - \alpha)$ may be less than unity (the anti A-J case).

Observe that if u is to have finite mean and variance, then $\lambda > 2$. As long as $\alpha > 1/2$, this is compatible with either the A-J or the anti A-J case.

IV. Multiplicative Uncertainty Reconsidered

I now turn to my critique of Peles and Stein's anti A-J theorem for multiplicative

[1]A formula for expected profit similar to that used in Section I may be obtained by writing $G = 1.G$ and integrating by parts.

[2]Fix s and define $\tilde{K}$ by $s\tilde{K}^{1-\alpha} = 1$. Then $H(K,s) = (s - i)K$ for $K \leqq \tilde{K}$, and $H(K,s) = K^\alpha(\lambda - 1)^{-1}(\lambda - (sK^{1-\alpha})^{1-\lambda}) - iK$ for $K \geqq \tilde{K}$. At $K = \tilde{K}$, both left and right derivatives of $H(K,s)$ with respect to K are equal to the positive number $(s - i)$. Hence $K_s > \tilde{K}$. Notice also that there is no shut-down option in this example. As long as $s > i$, we have $H(K,s) > 0$ for all sufficiently small K.

uncertainty. Throughout this section, it is assumed that the maximal quasi-rent function has the form

$$(13) \qquad R(K,u) = Q(K)\cdot(1 + u)$$

where $\mu = Eu$, and E is the expectation operator.[3]

As in Section III, let $H(K,s)$ denote expected net profit under regulation. Other notation is as in Section I. It is immediate from (13) that

$$(14) \qquad Q'(K_\infty)\cdot(1 + \mu) = i$$

Now consider the statement by Peles and Stein quoted in my introduction. There are two possible interpretations of this statement:

1) if s is sufficiently close to i, then $dK_s/ds > 0$;

2) if s is sufficiently close to i, then $K_s < K_\infty$.

Peles and Stein do not attempt a proof of 1), and it can be seen from the example of Section III that 1) is not in general correct. Peles and Stein do, however, attempt a proof of 2) on page 784 of their article.

To evaluate the Peles-Stein argument, let us first define some terms. Set

$$\theta = v(K_\infty, s)$$
$$\beta = (s - i)(1 - F(\theta))$$
$$\gamma = (1 + \mu)^{-1} \int_a^\theta (\mu - u) f(u)du$$

Using (14), one may show that

$$(15) \qquad H_1(K_\infty, s) = \beta - i\gamma$$

This equation is essentially equation (19) of the Peles-Stein article.

From the concavity assumption (2), we see that $H_1(K_\infty, s)$ has the same sign as the expression $(K_s - K_\infty)$. Thus from (15) $K_s - K_\infty \sim \beta - i\gamma$. Both β and γ are positive. Hence, as we already know, $(K_s - K_\infty)$ is indeterminate in sign.

Now consider the case where s is only slightly greater than i. Peles and Stein's argument for 2) may be translated into my notation as follows. As s approaches i from above, β approaches zero; but $\gamma > 0$; hence 2) holds.

This argument, however, ignores the fact that γ depends on θ and hence on s. If $v(K_\infty, i) = a$, then θ approaches a, and γ approaches zero, as s approaches i. If $i\gamma$ approaches zero faster than β, then 2) may fail to hold. It can be verified[4] that precisely this phenomenon occurs in the example of Section III when $(1 - \alpha)\lambda > 1$.

Evidently, a sufficient condition for 2) to be correct is that $v(K_\infty, i) > a$. By (13), and our definition of "critical state of nature" in Section I, we have $v(K_\infty, i) > a$ if and only if $Q(K_\infty)\cdot(1 + a) < iK_\infty$.

Thus, if there is positive probability of a state of nature sufficiently "bad" that the *unregulated* monopolist would make a loss at *his* optimal capital stock and that bad state of nature, then 2) will hold. This is yet another illustration of our point that a high degree of "noise" in the environment is favorable to anti A-J results.

[3]Peles and Stein assume that the expected value of u is zero. This assumption may be made without loss of generality by suitably redefining Q and u. Notice, however, that this normalization was not adopted in the example of Section III of this paper; in the interest of notational consistency, it is not adopted in Section IV.

[4]In this example, $a = 0$. For μ to be finite, we must have $\lambda > 1$, in which case $1 + \mu = \lambda/(\lambda - 1)$. Thus

$$K_\infty^{1-\alpha} = \frac{\alpha\lambda}{i(\lambda - 1)}$$

But then $\theta = \max(0, \theta_o)$, where θ_o is defined by the relation

$$1 + \theta_o = \frac{s\alpha\lambda}{i(\lambda - 1)}$$

Suppose that $(1 - \alpha)\lambda > 1$ and that s is sufficiently close to i that

$$1 < \frac{s}{i} < \frac{\lambda - 1}{\alpha\lambda}$$

Then $\theta_o < 0$, so $\theta_o = a = 0$, so $\gamma = 0$. Thus for all s sufficiently close to i, we have $\beta = s - i > 0$ and $i\gamma = 0$

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

Y. C. Peles and J. L. Stein, "The Effect of Rate of Return Regulation is Highly Sensitive to the Nature of the Uncertainty," *Amer. Econ. Rev.*, June 1976, *66*, 278–89.

On Regulation and Uncertainty: Reply

By YORAM C. PELES AND JEROME L. STEIN*

We are delighted that Nicholas Rau has attempted to generalize our result that: the effect of rate of return regulation is highly sensitive to the nature of the uncertainty. His paper has stimulated us to reflect further on generalizing and simplifying our joint results.

Originally, we stated the following results:

a) If uncertainty affects the maximal quasi-rents function $R(K,u)$ in a multiplicative way, $R(K,u) = R(K)(1+u)$, then a sufficiently large gap must exist between the regulated rate of return (s) and the cost of capital (i) to induce the firm to select *ex ante* a scale of plant greater than the scale chosen by the unregulated monopolist. The closer is the regulated rate to the cost of capital, the more likely is it that the regulated firm will select *ex ante* a smaller scale of plant than is chosen by the unregulated firm. Were that to occur, regulation would definitely be worse than no regulation.

b) If the uncertainty enters the maximal quasi-rents function in an additive way, $R(K,u) = R(K) + u$, then the conventional Harvey Averch and Leland Johnson (A-J) effect occurs (unless the regulation drives it out of business).

Rau's main conclusion is that: "... if the state of nature affects both the average and marginal return on capital, then whether an A-J or anti A-J effect prevails depends on the amount of randomness in the environment. The more 'noise,' the more likely is an anti A-J effect" (p. 190).

A general and simple statement of the regulation theorems is derived below which contains points a) and b) and Rau's conclusion as special cases.

*The Hebrew University, Jerusalem and Brown University, respectively.

I. A General Formulation of the Problem

A. *Expected Profit Maximization*

Let $R(K,u)$ be the maximal quasi rents that can be earned by an unregulated firm whose *ex ante* control is the scale of plant K. The exact way that uncertainty enters the $R(\cdot)$ function is specified below. The firm's *ex post* control is the labor input, chosen after the state of nature u is revealed. The *unregulated* firm selects scale of plant K_o such that expected profits π_o in equation (1) are maximized.

$$(1) \qquad \operatorname*{Max}_K \pi_o = \operatorname*{Max}_K \mathop{E}_u R(K,u) - iK$$

Under the usual assumptions, equation (1) implies that the expected marginal revenue product of capital $E_u R_K(\cdot)$ is equal to the cost of capital i, as stated in equation (1'). The optimum stock of capital chosen by the unregulated firm is $K = K_o$.

$$(1') \qquad \mathop{E}_u R_K(K_o,u) = i$$

The question is whether the regulated firm will select a capital stock greater than or less than K_o.

The regulated firm is not permitted to charge a price at which it earns more than sK of quasi rents; however, the scale of plant is determined before the state of nature is known. Partition the state of nature u into two disjoint nonempty sets S and T:

$$(2a) \qquad S \equiv [u \mid R(K,u) < sK]$$

$$(2b) \qquad T \equiv [u \mid R(K,u) \geq sK]$$

When the state of nature is in S, then the regulatory constraint is not effective. Given the scale of plant chosen before the state of nature is revealed, maximal quasi rents are less than sK when the state of nature is revealed. Point M'' (in Figure 1 of our orig-

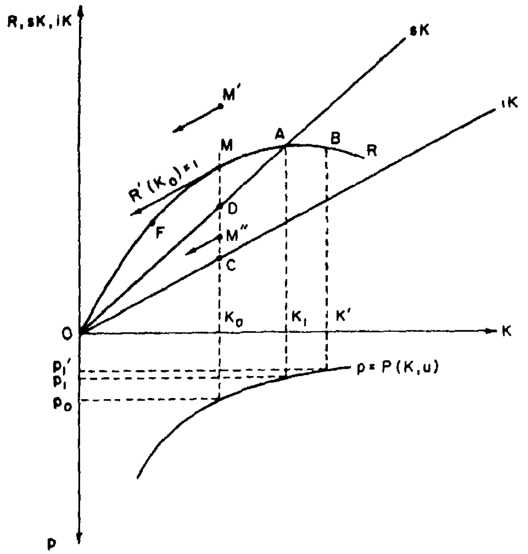


FIGURE 1

inal paper) is in set S. Figure 1 is reproduced here.

When the state of nature is in T, then regulation is effective. Points M and M' in Figure 1 are in set T. The firm is not permitted to earn quasi rents in excess of sK. The price will be lowered so that the firm just earns sK of quasi rents, thereby driving a firm with capital stock K_o to point D.

The regulated firm will select a scale of plant K_1 to maximize expected profits π_1 as defined in equation (3).

$$(3)\quad \underset{K}{\text{Max}}\ \pi_1 = \underset{K}{\text{Max}}\ [\underset{u \in S}{E} R(K,u) + sKPr(u \in T)] - iK$$

The expected maximal quasi rents, the terms in brackets, are a combination of quasi rents $R(K,u) < sK$ when u is in set S, and quasi rents sK when u is in set T. The latter occurs with probability $Pr(u \in T)$.

If the firm is to enter this business (i.e., if its *ex ante* control K is to be positive), the expected maximal quasi rents per unit of capital $E_u R(K,u)/K$ must exceed the cost of capital i. That means that the regulatory authority must select "s" so that equation (4) is satisfied for some $K > 0$.

$$(4)\quad \underset{u \in S}{E} R(K,u)/K + s\,Pr(u \in T) > i$$

The first term on the left-hand side is the expected quasi rents per unit of capital when $u \in S$; hence, it is less than s. In fact, it could very well be less than the cost of capital i. Graphically if $K = K_o$ then $E_{u \in S} R(K_o,u)$ is a weighted average of the points below the line sK, where the weights are the probabilities that $R(K_o,u)$ will lie in the interval $[0, sK_o)$.

Equation (4) vividly illustrates that it is erroneous to believe that the closer s is to i, the larger will be the chosen scale of plant. If the first term is less than i for all K then, as s is lowered towards i, there will come a point where the left-hand side is less than "i." When that occurs, no capital will be invested in this business. The regulatory authorities must guard against this possibility

Assume that s is chosen so that inequality (4) is satisfied. Determine the expected marginal revenue product of capital $ER_K(K,u)$ Differentiate expected quasi rents in equation (3) with respect to K and obtain equation (5).

$$(5)\quad \frac{d}{dK} \underset{u}{E} R(K,u) = \underset{u \in S}{E} R_K(K,u) + s\,Pr(u \in T) + \phi(K)$$

where the $\phi(K)$ term involves changes in the probabilities of u being in set S or T, as a result of variations in K. In our examples on pages 285 86 the $\phi(K)$ term was zero. Assume that $\phi(K)$ can be disregarded.[1]

The derivative of expected profits with respect to the scale of plant is given by equation (6), when the function is differentiable.

$$(6)\quad \frac{d\pi_1}{dK} = \underset{u \in S}{E} R_K(K,u) + s\,Pr(u \in T) - i$$

Evaluate $d\pi_1/dK$ at unregulated point K_o for two reasons. First, the expected

[1] Rau, in deriving his equation (5) from the previous equation, failed to differentiate the $v(K,s)$ term in the upper limit of integration. Implicitly, he must have assumed that $\partial v(K,s)/\partial K = 0$ as we do. Otherwise, his equation (5) is incorrect.

profit function is certainly differentiable at that point. In the certainty case, $d\pi_1/dK$ is $(s - i)$ at K_o, the unregulated capital stock.[2] Second, it provides a quick and easy answer to the question of whether an A-J or anti A-J effect will occur. If $d\pi_1/dK$ at $K = K_o$ is positive then the firm will select *ex ante* a scale of plant greater than K_o. If $d\pi_1/dK$ at $K = K_o$ is negative, then the firm will select *ex ante* a capital stock less than K_o.

B. *An Evaluation of* $E_{u \in S} R_K(K, u)$

The only difficult problem is to evaluate $E_{u \in S} R_K(K, u)$ in equation (6): The expected marginal product of capital when u is in set S. This is solved in the present section for more general functions than were considered in our earlier paper.

Consider $R_K(K, u)$ when the capital stock is equal to the unregulated scale K_o. Then $R_K(K_o, u)$ is a function of the disturbance term u which is assumed to have a zero expectation. Equation (7) is derived from a Taylor expansion around $Eu = 0$.

$$(7) \quad R_K(K_o, u) = R_K(K_o, 0) + R_{Ku}(K_o, 0)u$$

According to equation (1′), the expectation of the marginal product of capital over the *entire range* of u, $E_u R_K(K_o, u)$ is equal to the cost of capital i. According to equation (7), the expectation $E_u R_K(K, u)$ over the entire set of u is $R_K(K_o, 0)$. When these results are combined:

$$(8) \quad \underset{u}{E}\, R_K(K_o, u) = R_K(K_o, 0) = i$$

Substitute (8) into (7) and obtain

$$(9) \quad R_K(K_o, u) = i + R_{Ku}(K_o, 0)u$$

The expected marginal product of capital in equation (5) is equation (10), since we use equation (9).

$$(10) \quad \underset{u}{E}\, R_K(K_o, u) = \underset{u \in S}{E}\,[i + R_{Ku}(K_o, 0)u] + s\,Pr(u \in T)$$

$$= i\,Pr(u \in S) + R_{Ku}(K_o, 0)\,E[u \mid u \in S] + s\,Pr(u \in T)$$

Since $Pr(u \in S) + Pr(u \in T) = 1$, write equation (10) as

$$(11) \quad \underset{u}{E}\, R_K(K_o, u) = i + (s - i)Pr(u \in T) + R_{Ku}(K_o, 0)E[u \mid u \in S]$$

Substitute equation (11) into $d\pi_1/dK = E_u R_K(K_o, u) - i$ and derive equation (12) or (12′) for the general case of the regulated firm. It is marginal profit for the regulated firm, evaluated at the unregulated scale K_o.

$$(12) \quad \left.\frac{d\pi}{dK}\right|_{K_o} = i\,Pr(u \in S) + R_{Ku}(K_o, 0)\,E[u \mid u \in S] + s\,Pr(u \in T) - i$$

or

$$(12') \quad \left.\frac{d\pi}{dK}\right|_{K_o} = (s - i)\underset{(+)}{Pr}(u \in T) + R_{Ku}(K_o, 0)\underset{(-)}{E}[u \mid u \in S]$$

This most general equation contains the Peles-Stein and Rau results as special cases.

C. *When the Anti A-J Effect Occurs*

At the unregulated point, monopoly quasi rents $r_o K_o$ exceed $sK_o > iK_o$. This means that $(s - i)Pr(u \in T)$ is positive. Component $E[u \mid u \in S]$ in equation (12) or (12′) must be negative, because a negative value of u (equal to MD in Figure 1) is required to reduce monopoly quasi rents to sK_o.

$$(13) \quad E[u \mid u \in S] = -\gamma < 0$$

Obviously, the first term in equation (12′), $(s - i)Pr(u \in T)$, is responsible for the A-J effect. An anti A-J effect can only occur if a negative second term, $-R_{Ku} \cdot (K_o, 0)\gamma$, outweighs the positive effect.

[2] In the certainty case, maximal quasi-rents function is *ODABR* in Figure 1 of our original paper. This function is *not* differentiable at point *A*, the A-J point. Hence, in the certainty case, the optimal solution for a regulated monopolist does *not* occur where $d\pi_1/dK$ is zero. For this reason, we use the unregulated point K_o as our frame of reference: π_1 is differentiable at $K = K_o$.

If $R_{Ku}(K_o,0) = 0$, then the A-J effect will occur. This condition is precisely the additive case $R(K,u) = R(K) + u$ discussed in our paper.[3] The value of the marginal product of capital is independent of the disturbance term in this additive case. This indeed was the explanation we offered on page 285 of our paper.

Our multiplicative case, $R(K,u) = R(K)(1 + u)$, is a special case of the condition $R_{Ku}(K_o,0) > 0$, which is sufficient for our anti A-J effect to occur.[4] In this multiplicative case, the expected marginal product of capital

$$\underset{u}{E}\, R_K(K_o,u) = i + (s - i)Pr(u \in T) - \gamma i$$

is affected by the "noisiness" of the system, which must be associated with the magnitude of $\gamma > 0$.

In the general case of equation (12′), we derive the following anti A-J proposition.

PROPOSITION: *If* $R_{Ku}(K_o,0)$ *is positive then:* (a) *the smaller* $(s - i)$ *and* (b) *the "noisier" is the system, i.e., the greater is* γ *then the more likely is it that the regulated scale of plant* K_1 *will be less than that selected by the unregulated monopolist.*

II. An Intuitive Explanation of the Proposition

An intuitive explanation of our general proposition will clarify the relation between our original paper and Rau's. Consider the optimum unregulated scale of plant K_o (Figure 1). Table 1 contains the average quasi rents per unit of capital $R(K)/K$ and the marginal product of capital $R'(K)$, in the deterministic case. When there is regulation, the effective average quasi rents per unit of capital that the firm is allowed to earn is constant at s (for $K < K_1$ in Figure 1). Since the average return to capital is constant, the effective marginal product of capital is also equal to s. When there is uncertainty, the situation is described by Table 2. Again, the functions are evaluated at the unregulated scale K_o.

TABLE 1—DETERMINISTIC CASE: $K = K_o$

	Average $R(K)/K$	Marginal $R'(K)$
Unregulated	r_o	i
Regulated	s	s

The expected marginal product of capital at $K = K_o$ is i, when there is no regulation. With regulation, the expected marginal product of capital at $K = K_o$ is a linear combination of the marginal product of capital for various values of the stochastic variable u. Suppose that u assumed values $(u', 0, u'')$ that produced maximal quasi rents (M', M, M'') in Figure 1. Insofar as the maximal quasi-rents function passes through M' or M, then $R(K) > sK$. This means that effective average quasi rents per unit of capital that the firm is allowed to earn are s. Marginal quasi rents are also s, under those conditions. The marginal quasi rents s occur with probability $Pr(u \in T)$.

If $u = u''$ so that the maximal quasi-rents function passes through M'', then the slope of the maximal quasi-rents function may be the same as it was at M or it may be adversely affected by the decline in u to u''. If $R_{Ku}(K_o,0) > 0$, then the slope of the maximal quasi-rents function that passes through M'' will be less than its slope at M (equal to i). That is, a negative value of u lowers *both* the *average* and the *marginal* quasi-rents function when $R_{Ku}(K_o,0) > 0$.

When that occurs, the expected marginal product of capital at $K = K_o$ is a linear combination of s (when points M or M' occur) and *something less than* i (when point M'' occurs). If s is reduced towards i, then the expected marginal product of capital will be reduced below i. That is the anti A-J effect, and it is exactly what is contained in Rau's conclusion cited above.

[3] (i) $\frac{\partial^2 R}{\partial K \partial u} = R_{Ku}(K,u) = 0$
which implies (ii) $\frac{\partial R}{\partial K} = R'(K)$
which implies (iii) $R(K,u) = R(K) + u$
or a function of u. This is the additive case.

[4] From (i) $R(K,u) = R(K)(1 + u)$
derive (ii) $R_K(K,u) = R'(K)(1 + u)$
which implies (iii) $R_{Ku}(K,u) = R'(K)$
At $K = K_o$: (iv) $R_{Ku}(K_o,u) = R'(K_o) = i > 0$

TABLE 2—UNCERTAINTY CASE: $K = K_o$

	Average	Marginal
Unregulated	r_o	i
Regulated	$s\,Pr(u \in T) + \int_{u \in S} \frac{R(K)}{K} g(u)\,du < s$	$i\,Pr(u \in S) + s\,Pr(u \in T) - \gamma R_{Ku}(K_o, 0)$

If variations in u shift the maximal quasi-rents function $OMABR$ (Figure 1) parallel to itself, then the marginal product of capital at any given K is unaffected by the value of u: i.e., $R_{Ku}(\cdot) = 0$ in this case. Then, the slope of the maximal quasi-rents function is the same at M'' as it is at M or M', equal to i. Consequently, the weighted average marginal product of capital at points (M', M, M'') is a linear combination of s and i, which exceeds i. That produces the A-J effect.

The crucial question for the existence of the A-J effect or the anti A-J effect is whether shifts in the average quasi-rents function, resulting from variations in u, affect the marginal as well as the average product of capital. We think that this analysis generalizes both our results and those of Rau.

REFERENCES

Y. Peles, and J. L. Stein, "The Effect of Rate of Return Regulation is Highly Sensitive to the Nature of the Uncertainty," *Amer. Econ. Rev*, June 1976, *66*, 278–89.

Nicholas Rau, "On Regulation and Uncertainty: Comment," *Amer. Econ. Rev.*, Mar. 1979, *69*, 190–94.

The Supply of Storage: Stein vs. Snape

By BARRY A. GOSS*

Jerome Stein (1961, 1964) published a portfolio selection model of individual discretionary hedger decision making in a futures trading context, and used this model as a basis for his theory of the simultaneous determination of spot and futures prices. While this model has several limitations (see for example the author and Basil Yamey), it is the purpose of this note to show that it does not have the deficiency attributed to it by Richard H. Snape, who argued that Stein's neglected substitution effect, when accounted for, gives rise to the possibility of an unstable storage market equilibrium.

Stein's model determines the proportion of stock to be hedged for a risk-averse individual whose assumed aim is maximization of expected utility. The expected return equations are

(1) $u = p^* - p - m$

(2) $h = (p^* - p - m) + (q - q^*)$

where u = expected return per unit on unhedged stock
h = expected return per unit on hedged stock
p = current spot price
q = current futures price
p^* = expected spot price
q^* = expected futures price
m = marginal carrying cost (net)

Using the variance of expected return as a measure of risk, the risk per unit of unhedged stock is

$$V(u) = V(p^*)$$

while the risk per unit of hedged stock is

$$V(h) = V(p^*) + V(q^*) - 2\,cov\,p^*q^*$$

Assuming that hedged stock has a smaller risk and smaller expected return per unit than unhedged stock, there is a direct relationship between risk and expected return in the opportunity locus, as the proportion of stock unhedged varies between zero and one. In Stein's 1961 paper this relationship was assumed to be linear, but in his 1964 paper it was shown to have a negative second derivative. In the quadratic expected-utility function the assumed diminishing marginal utility of income gives a rising marginal rate of substitution. The optimum proportion of stock to be hedged by the individual is that which maximizes expected utility.

The model does indeed have several limitations. It does not precisely determine the individual's spot market position, making it simply an increasing function of maximum expected utility; nor does it permit an individual to overhedge, or to be long in spot and long in futures, positions which a discretionary hedger may wish to take if he believes that spot and futures prices will fall, or rise, respectively. Moreover, there is the difficulty that the variance is not an efficient measure of risk when spot and futures prices change *ceteris paribus* (see the author and Yamey).

Stein's model does not however, have the deficiency attributed to it by Snape who argued that Stein does not allow "fully" for the substitution effect between hedged and unhedged stocks in his model of intertemporal price determination (see Snape, p. 220). Snape says that if the current futures price rises *ceteris paribus* there will be an increase in total demand for stocks, and in the proportion of hedged stock, but the quantity of unhedged stock may rise or fall; and "... if p fell *ceteris paribus*... because of the change in relative profitability, demand for one of the components could fall. As a consequence (Stein's) SS curve ... could be negatively sloped" (p. 220). This conclusion is based on a misinterpretation of Stein's model.

It is clear from the model of individual

*Monash University.

decision making, that for a rise in the futures price, *ceteris paribus*, the substitution effect favors hedged stock, is opposite in sign to the income effect, and is dominant. Stein says "As the slope of the transformation line ... is decreased there is a substitution effect. The ratio of unhedged to total stock will be decreased as the hedging of stock becomes relatively more attractive.... The substitution effect will be dominant ... the proof of this proposition is found in Tobin [5, p. 79]" (1961, p. 1016). In Figure 1, with individual equilibrium initially at P, a rise in the futures price *ceteris paribus* causes the opportunity locus to pivot on U to become $H'U$, with the consequent new equilibrium at Q. On the argument the optimum proportion of stock unhedged will decline from OA to OB. The quantity of hedged stock must rise, the quantity of unhedged stock may go either way, as Snape observes. Indeed, Snape does not seem to quarrel with this analysis of individual equilibrium.

For a fall in the spot price *ceteris paribus*, u and h are affected equally, there is no change in relative profitability, contrary to Snape, and the income effect favors the proportion of unhedged stock (on the assumptions indifference curves become flatter as they cross an ordinate at successively

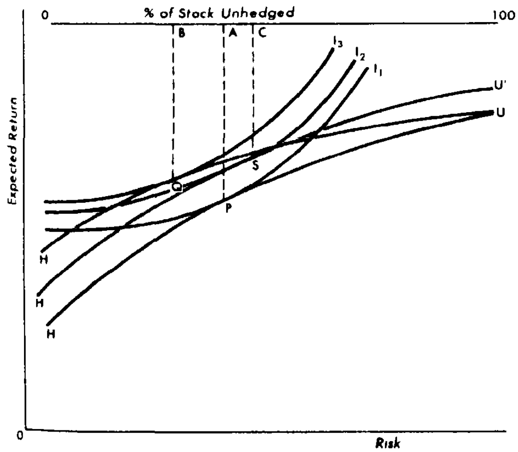


FIGURE 1. (BASED ON STEIN, 1961, WITH MODIFICATION REQUIRED BY STEIN, 1964)

higher levels of expected return). In Figure 1, with individual equilibrium initially at P, a fall in the spot price *ceteris paribus* causes the opportunity locus to undergo a parallel shift from HU to $H''U''$ with the new equilibrium at S, which on the argument must lie to the right of P. The optimum proportion of stock unhedged therefore increases from OA to OC. The quantity of unhedged stock must increase, and the quantity of hedged stock, as Snape observes, may go either way.

Stein writes the market demand for stocks (supply of storage) as

$$(3) \quad S_D = U(p^* - p - m) + H(p^* - q^* + b - m)$$

where $b = q - p$, and p^*, q^* are redefined as the weighted averages of individuals' expectations of spot and futures prices, respectively. In (3) Stein makes $U' > 0$, $H' > 0$, and calls U the market demand for unhedged stock and H the market demand for hedged stock. (See Stein, 1961, p. 1017.) It is evidently this analysis with which Snape disagrees, objecting to U', $H' > 0$, on the ground that "demand for one of the components could fall" (p. 220). This criticism is not valid because U' and H' are the partial derivatives

$$\frac{\partial U}{\partial(p^* - p - m)}; \quad \frac{\partial H}{\partial(p^* - q^* + b - m)}$$

$U' > 0$ is interpreted to mean that the demand for unhedged stocks rises with u, *ceteris paribus*, while $H' > 0$ is interpreted to mean that the demand for hedged stocks rises with h, *ceteris paribus*. These results are quite consistent with the preceding model of individual decision making. Each of the partial derivatives U', H' is dependent upon a different set of *ceteris paribus* conditions. $U' > 0$ does not imply that the demand for *hedged* stocks rises with u; nor does $H' > 0$ imply that the demand for unhedged stocks rises with h.

The supply of stocks (i.e., demand for storage) is written as $S_{-1} + X(p,a)$, where S_{-1} is opening stock, X is excess supply of current production, and a is a parameter.

Hence the condition for equilibrium in the storage market is

$$(4) \quad U(p^* - p - m) + H(p^* - q^* + b - m) = S_{-1} + X(p,a)$$

The (b,p) locus for equilibrium in this market (SS) is found by differentiating (4) with respect to p and solving for $\partial b/\partial p$, giving

$$U_p + H_b \frac{\partial b}{\partial p} = X_p$$

whence

$$(5) \quad \frac{\partial b}{\partial p} = \frac{X_p - U_p}{H_b}$$

where subscripts refer to partial derivatives. Clearly $U_p < 0, H_b > 0, X_p > 0$, so that $\partial b/\partial p > 0$. There is no ground for a negatively sloped SS curve in Stein's model.

Unique equilibrium values for spot and futures prices are found by deriving the (b,p) locus for equilibrium in the futures market (FF) and seeking the coincidence of these two loci. In the futures market the supply of futures is assumed to be equal to the supply of hedged storage, and is given by the second term on the left-hand side of equation (4). The demand for futures contracts is assumed to be provided by speculators in futures, their purchases being inversely related to the current futures price, *ceteris paribus*. By an argument similar to that yielding equation (5) above it is shown that the locus FF, which is not the subject of dispute between Snape and Stein, is negatively sloped. Intersection of these loci yields equilibrium values for the spot price p_0 and the price spread b_0: hence the equilibrium futures price $q_0(\equiv b_0 + p_0)$ is also determined (see Figure 2).

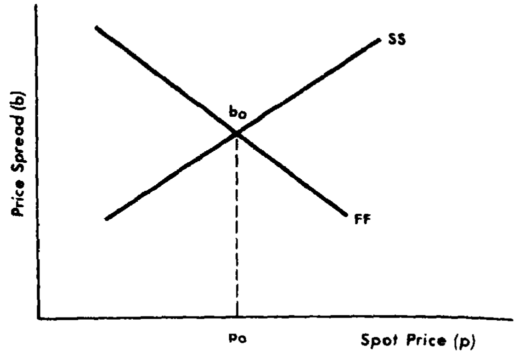


FIGURE 2. (BASED ON STEIN, 1961)

Without reproducing Stein's (1961) comparative statics (pp. 1021–24) it would appear that the stability of this final equilibrium remains unchallenged, there being no other alleged ground for instability apart from a negatively sloped SS curve, a suggestion we have shown to be unfounded. This is not to deny the possibility of unstable equilibrium in forward markets, which Snape wished to discuss. The analysis discussed in this note, however, is not the basis for it.

REFERENCES

Barry A. Goss and Basil S. Yamey, "Introduction," in their *The Economics of Futures Trading*, London 1976.

R. H. Snape, "Forward Exchange and Futures Markets," in Ian A. McDougall and R. H. Snape, eds., *Studies in International Economics, Monash Conference Papers*, Amsterdam 1970.

J. L. Stein, "The Simultaneous Determination of Spot and Futures Prices," *Amer. Econ. Rev.*, Dec. 1961, *51*, 1012–25.

———, "The Opportunity Locus in a Hedging Decision: A Correction," *Amer. Econ. Rev.*, Sept. 1964, *54*, 762–63.

J. Tobin, "Liquidity Preference as Behavior Toward Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–86.

Labor Supply under Uncertainty: Note

By GIDEON YANIV*

In a recent note in this *Review*, David Sjoquist corrects some of the results obtained earlier by Michael Hartley and Nagesh Revankar (H-R). In their model of an individual's labor supply decision, uncertainty arises because of possible unemployment. Sjoquist argues that under their assumption of a utility function with no certainty bias, their method of maximizing utility of expected values should yield the same results as maximizing expected utility. He sets up the individual's problem as

$$(1)\quad \max_N EU = (1-\pi) \cdot U(wN + Y, T - N) + \pi U(u + Y, T)$$

where EU is expected utility (satisfying the usual concavity assumptions), π the probability of being unemployed (unemployment rate), w the wage rate, N the hours of labor supplied, Y nonlabor income, T total time, and u the unemployment insurance payments. Since it is obvious from (1) that the individual's choice of N^* does not affect his utility from the unemployment state, Sjoquist asserts that $dN^*/d\pi = 0$, whereas H-R conclude that $dN^*/d\pi > 0$ if the income effect is negative but very small.[1]

In an attempt to reach a more convincing result from his expected-utility approach, Sjoquist suggests that hours of effort supplied but not worked are spent by the individual in search of a position, and can thus be regarded as hours of actual work, except that the wage rate is zero. Under this assumption the individual's problem is stated as

$$(2)\quad \max_N EU = (1-\pi) \cdot U(wN + Y, T - N) + \pi U(u + Y, T - N)$$

so that by taking the first-order condition and differentiating it with respect to π, Sjoquist shows that $dN^*/d\pi < 0$ regardless of the individual's attitude toward risk.

Sjoquist's assumption, however, is hardly reasonable. As he himself seems to realize, hours allocated to work when employed are not necessarily identical to hours allocated to search when unemployed. More important, once in a state of unemployment, the individual is not provided with any probabilities of finding work by entering the search process. Why, then, would he spend any time at all looking for a job he surely will not get? Indeed, Sjoquist's unappealing conclusion that increases in the rate of unemployment cause individuals to supply less labor follows directly from the unrewarding role he attributes to labor in the unemployment state.

The desired dependency of labor supply on the unemployment rate can nevertheless be achieved by assuming that unemployment insurance payments are proportional, with a given rate of $0 < u < 1$, to the loss of earnings. This is, in fact, a common provision in many unemployment compensation programs, where past earnings usually serve as a measure of current loss.[2] Within the present framework, however, the individual can be assumed to apply to an official employment bureau, where he declares his willingness to work N^* hours in the given period. The bureau has a probability of $1 - \pi$ of finding him a suitable job at which he will earn wN^* units of income and

*Hebrew University and Bureau of Research and Planning, National Insurance Institute, Jerusalem. I wish to thank Giora Hanoch, Marjorie Honig, Joram Mayshar, Nehemia Shiff, Yossi Tamir, and a referee for helpful comments.

[1]The discrepancy in the two results is not explained by Sjoquist. The discussion which follows is not related to this issue, however.

[2]In the United States, for example, the compensation rate is roughly 50 percent of past earnings. In Israel, the percent varies from 30 to 70 percent, decreasing as past earnings increase (see U.S. Department of Health, Education and Welfare).

enjoy $T - N^*$ hours of leisure. If, with a probability of π, the bureau fails to provide him with such a job, he will receive uwN^* units of income as unemployment insurance payments, having to devote his entire time T to leisure activities.[3]

Formally, the individual's maximization problem now becomes

$$(3)\quad \max_N EU = (1-\pi) \cdot U(wN + Y, T - N) + \pi U(uwN + Y, T)$$

yielding as a first-order condition

$$(4)\quad \frac{dEU}{dN} = 0 = (1-\pi)(wU_1^A - U_2^A) + \pi uwU_1^B$$

where A and B denote the states of employment and unemployment, respectively. The second-order condition, which is assumed negative, is obtained by differentiating (4) with respect to N:

$$(5)\quad \frac{d^2EU}{dN^2} \equiv D = (1-\pi)(w^2U_{11}^A - 2wU_{12}^A + U_{22}^A) + \pi(uw)^2U_{11}^B < 0$$

Differentiating now (4) with respect to π yields

$$(6)\quad \frac{dN^*}{d\pi}D = wU_1^A - U_2^A - uwU_1^B$$

or by substituting (4) into (6):

$$(7)\quad \frac{dN^*}{d\pi} = -\frac{uwU_1^B}{(1-\pi)D} > 0$$

That is, the individual's labor supply is positively sensitive to changes in the unemployment rate. By offering more hours of work when the unemployment rate increases, the individual in fact insures himself against the realization of unemployment.

[3]Note that the present assumption rules out the possibility that an individual will receive unemployment payments even if he willingly avoids the labor market (declares $N^* = 0$). A deliberate refusal of available jobs by an individual who has initially declared $N^* > 0$ can be made unattractive by rejecting his claim for compensation.

It seems only natural now to examine how a change in the level of the unemployment compensation affects the individual's supply of labor. Whereas H-R conclude that $dN^*/du < 0$ if leisure is a normal good (regarding u as a flat rate benefit), formulation (1) trivially yields $dN^*/du = 0$. Formulation (2), adopted by Sjoquist, can be easily shown to yield $dN^*/du = \pi U_{21}^B/\overline{D}$, where $\overline{D}$ is the appropriate second-order derivative of expected utility with respect to N. To obtain in this case $dN^*/du < 0$, a stronger requirement than normality ($U_{21}^B > 0$) is needed. However, differentiating the first-order condition of formulation (3) with respect to the rate of unemployment insurance yields

$$(8)\quad \frac{dN^*}{du} = -\frac{\pi w(U_1^B + uwNU_{11}^B)}{D}$$

so that substitution and income effects should be evaluated to determine the sign of (8). Assuming, however, that there is no income from nonlabor sources ($Y = 0$), we can rewrite (8) as

$$(9)\quad \frac{dN^*}{du} = -\frac{\pi wU_1^B}{D}[1 - R_R(I^B)]$$

where I^B denotes total insurance payments in the state of unemployment and $R_R(I) = -IU_{11}(I)/U_1(I)$ is the Arrow-Pratt relative risk-aversion measure with respect to money income.

Hence, the way in which the individual's labor supply responds to changes in the rate of the compensation payments depends upon the nature of his risk aversion. There seems to be a general presumption that relative risk aversion is a nondecreasing function of income. We may conclude that if $R_R(I^B)$ does not vary with income then for any value of u:

$$(10)\quad R_R(I^B) \gtreqless 1 \Longleftrightarrow \frac{dN^*}{du} \lesseqgtr 0$$

However, if $R_R(I^B)$ rises with income then $dN^*/du > 0$ might hold for sufficiently low

values of u, while increases in u may eventually lead to a change in sign.

REFERENCES

M. J. Hartley and N. S. Revankar, "Labor Supply Under Uncertainty and the Rate of Unemployment," *Amer. Econ. Rev.*, Mar. 1974, *64*, 170–75.

D. L. Sjoquist, "Labor Supply Under Uncertainty: Note," *Amer. Econ. Rev.*, Dec. 1976, *66*, 929–30.

U.S. Department of Health, Education and Welfare, *Social Security Programs Throughout the World*, res. rept. no. 48, Washington 1975.

Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South: A Reply to Fogel-Engerman

By THOMAS L. HASKELL*

Robert Fogel and Stanley Engerman's (hereafter, F-E) recent article on the relative efficiency of slave agriculture is intended to be a defense of *Time on the Cross*, yet in it the authors silently abandon the central theme of that book. From the outset what seemed most dubious to me and to other critics was F-E's use of the efficiency calculation as a basis for quite literal-minded inferences about the quality of plantation labor.[1] In their latest statement I find far less to disagree with because the dubious inferences have disappeared. If we are entitled to judge from what they now *omit* to say, Fogel and Engerman have changed their minds and thereby silently concede a great deal to their critics.

The crux of the matter is the source of efficiency. In both their book and their recent article, F-E locate one major source in the foresight, rationality, and energy with which plantation managers organized and disciplined their work force. But in the book the authors were equally enthusiastic—more so, in fact—about the slave's own personal contribution to plantation efficiency. The authors originally attributed the superior efficiency of plantations to

> ... the *combination* of the superior management of planters and the superior quality of black labor... In a certain sense, all, or nearly all of the advantage is attributable to the high quality of slave labor, for the main thrust of management was directed at improving the quality of labor. How much of the success of the effort was due to the management, and how much to the responsiveness of workers is an imperative question, but its resolution lies beyond the range of current techniques and available data.
> [1974, p. 210, emphasis added]

Beyond resolution or not, the authors seemed in their book consistently to resolve the question in favor of the slaves' contribution. One of their principal aims was to stamp out the "myth of black incompetence" (p. 223) and erase the "false stereotype of black labor" (p. 215) that abolitionists, ironically, had done so much to create. To these tasks they devoted a major part of the book. The average slave field hand, they said, was "harder working and more efficient than his white counterpart" (p. 5). They pointed to their efficiency calculation to prove that slaves typically were "diligent and efficient workers" (p. 263) and that slaves were "imbued like their masters with a Protestant ethic" (p. 231). In sum, the efficiency calculation served as the foundation for inferences about the admirable performance of slave laborers that formed what the authors themselves regarded as the central theme of their book—"the record of black achievement under adversity" (p. 264).

There is hardly a hint of all this in their latest statement. They now regard slave labor as efficient only in the production of four crops. This is a remarkably fragile sort of "efficiency" that cannot be translated into general farming, much less into factory

*Rice University.

[1]See the author. In the Paul David et al. volume, Paul David and Peter Temin noted that F-E used the term "efficiency" in a way that carried "... strong *physical* or technical connotations. Indeed it is precisely this physical interpretation upon which Fogel and Engerman rely in arguing that they can infer something about the 'task' efficiency of slaves compared with free laborers on the basis of the comparative rate at which a given bundle of labor, land, and capital inputs could be transformed into agricultural 'output' by slave and free farms" (p. 281). David and Temin concluded that "Fogel and Engerman's factor productivity measures can at best speak to the issue of the comparative 'revenue-earning efficiency' of the southern agricultural system ..." (p. 224 25). Similarly, Gavin Wright concluded that "... the regional index of total-factor-productivity is not a meaningful basis for drawing inferences about the relative hard-workingness of slave and free labor ..." (1976, p. 313).

occupations.[2] Moreover, the source of even this frail advantage appears now to lie in the master and in the impersonal system he designed. The slave is now merely a cog in a machine, and his contribution has shrunken into an unrecognizable vestige of its former richness: he now supplies the single undifferentiated quality of "intensity"—evoked, we are allowed to assume, more by the whip than the Protestant ethic. Notice also that the authors' new claim, that slaves worked so intensely that they accomplished in thirty-five minutes what it took free farmers an hour to do, is merely a logical deduction from the efficiency calculation. It is not supported by independent evidence.[3]

Efficiency is now said by Fogel and Engerman (1977) to stem from "... the persistence with which planters sought to exploit complementarities and interdependencies ..." (p. 290); from an elaborate division of labor that concentrated individual performance and generated "an assembly line type of pressure" (p. 291) between various task groups; from "time-motion studies" (p. 291) and production quotas; from the planter's careful selection of crops whose seasonal labor demands were compatible; from the forced labor of pregnant women and new mothers, who were kept at work in the fields as long as possible. Conspicuous by its absence from this account is the image of the achievement-oriented slave of *Time on the Cross*, who threw himself into work because his interests converged with those of his master.

Fogel and Engerman candidly admit now that "It was the system ... that made slave laborers more efficient than free laborers"—and they add: "... on farms that specialized in certain crops" (1977, p. 294). Carefully trimmed and tailored though it is, this newly formulated claim for the efficiency of southern agriculture has important ramifications and deserves careful consideration. Some specialists probably will remain unconvinced. But for my own purposes it is immaterial whether or not the South enjoyed a substantial advantage in a kind of "efficiency" that pertained only to four crops and stemmed from an impersonal system. *The critical point is that efficiency in this narrow and technical sense will not sustain the argument of which it was originally the foundation.* By giving up their attempt to portray the plantation experience as a fitting object of modern black pride—not only because blacks withstood it, but also because they felt they had a strong stake in its success—F-E have abandoned the central theme of *Time on the Cross* and the very claim that brought that book to the attention of the general reading public.[4]

[2]See F-E, 1977, pp. 292–93.
[3]See F-E, 1977, p. 293.

[4]I must also protest F-E's statement that I, among others, argued that "... the high relative efficiency of southern agriculture ... is merely an artifact of the temporarily inflated price of cotton in 1860 ..." (1977, p. 280). Neither I nor anyone else argued merely this. Fogel and Engerman's relative price ratios do not respond to my original criticism any better now than when they were first made public at the Rochester Conference on *Time on the Cross* in 1974. Subsequent econometric work by Wright and others appears to me to have validated my original point about cotton demand, but since that work also supercedes my own, both in expertise and weight of evidence, I happily leave the question to those best qualified to debate it. See Wright's contributions to this discussion.

REFERENCES

Paul A. David et al., *Reckoning with Slavery: A Critical Study in the Quantitative History of American Negro Slavery*, New York 1976.

Robert W. Fogel and Stanley L. Engerman, *Time on the Cross*, Vol. I, Boston 1974.

——— **and** ———, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South," *Amer. Econ. Rev.*, June 1977, *67*, 275–96.

T. L. Haskell, "Were Slaves More Efficient? Some Doubts About *Time on the Cross*," *New York Rev. of Books*, Sept. 19, 1974, 38–42.

G. Wright, "Slavery and the Cotton Boom," *Explor. Econ. Hist.*, Oct. 1975, *12*, 439–51.

———, "Prosperity, Progress and American Slavery," in Paul A. David, et al., *Reckoning with Slavery*, New York 1976.

The Relative Efficiency of Slave Agriculture: A Comment

By DONALD F. SCHAEFER AND MARK D. SCHMITZ*

In a recent article in this *Review*, Robert Fogel and Stanley Engerman (hereafter, F-E) defended their previous assertion that pre-Civil War southern agriculture was more efficient than its northern counterpart. The new study rebuked a number of critiques of their prior works (1971, 1974) and concluded that incorporation of suggested revisions actually widened the estimated differential in total factor productivity to 39 percent. The study also reiterated their finding that total factor productivity increased with the size of slaveholding. Using the Parker-Gallman cotton South sample for 1860, F-E estimated that farms with over fifty slaves enjoyed a level of efficiency over one-third higher than farms with no slaves.[1]

The objective of this note is to demonstrate that F-E have misinterpreted the latter finding. They viewed the positive association between size of slaveholding and efficiency as an indicator of the proficiency of slave labor. In addition, they argued that this association indicated the existence of scale economies in production but that these economies only existed beyond some minimum slaveholding where the regimented gang labor system could be used. It is our contention that in making these assertions, F-E combined and confused two concepts—the number of slaves and scale. The latter term is generally defined in terms of output or the general level of inputs, while the former only provides an imperfect index of one input, labor. The two concepts are not necessarily consistent measures of size and it is therefore necessary to distinguish between them in the type of analysis that F-E pursued. The preferred distinction between the concepts implies that there are actually two issues to appraise. The first—the F-E hypothesis—is whether larger slaveholdings resulted in higher productivity and the second is whether scale (as correctly measured by output) caused higher efficiency. Obviously, the two issues are not mutually exclusive, but it is our contention that F-E's failure to separate the issues led them to draw invalid inferences concerning the relationship between slavery and agricultural efficiency.

In the analysis that follows we have employed the same data as F-E; namely, the Parker-Gallman sample for the cotton South in 1860. From this sample of 5,228 farms, we have excluded the same 929 farms as F-E so that we are dealing with exactly the same subsample. Further, we have defined land, labor, capital, and output using the measures put forward by F-E; in other words, the data in this paper are identical to those of F-E.

I. Slaveholding and Productivity

Fogel and Engerman define total factor productivity (G) for slaveholding class s as

$$(1)\qquad G_s = \frac{\Sigma O_{si}}{(\Sigma L_{si})^{\alpha_L}(\Sigma K_{si})^{\alpha_K}(\Sigma T_{si})^{\alpha_T}}$$

where O, L, K, and T are output, labor, capital, and land, respectively; the α's are

*Departments of economics, North Carolina A & T State University; and University of Delaware and University of Washington, respectively. We would like to thank Robert Fogel and Stanley Engerman for providing detailed information on their input and output measures. The research was partially supported by a National Science Foundation Grant and funds from the Schools of Business and Economics at North Carolina A & T State University and the University of Delaware.

[1]Throughout this note we use the terms total factor productivity, productivity, and efficiency interchangeably. This is also true for the terms output and scale which are introduced below. For a description of the cotton South sample, see James Foust and Dale Swan, Appendix A. Information on this and other data sets relevant to the period can be found in Fogel and Engerman (1974, pp. 21-25).

TABLE 1—TOTAL AND PARTIAL PRODUCTIVITY INDEXES BY SLAVEHOLDERS: FOGEL-ENGERMAN ORIGINAL COTTON SOUTH RESULTS

Slave-holdings	Sample Size	Partial Productivity Indexes[a] Land	Labor	Capital	Total Factor-Productivity Index[a]
None	2081	100	100	100	100
1-15	1484	79.3	128.5	92.1	107.7
16-50	603	93.5	186.0	117.0	144.7
51 or more	131	69.6	197.5	91.4	133.5

Source: Parker-Gallman cotton South sample.

[a]Free Farms = 100 Output is net crop, pork, beef, and dairy outputs valued at 1860 prices. Land is the cash value of the farm. Labor is measured in prime male equivalents. Capital is the value of machinery and implements. Partial productivity is measured by the ratio of output to input Total factor productivity is from equation (1) with $\alpha_T = .25$, $\alpha_L = .58$, and $\alpha_K = .17$

Cobb-Douglas output elasticities; and i denotes the individual farms within the classes. Four slaveholding classes were considered and the F-E results for the cotton South are reproduced here as Table 1. Their estimates provided further documentation of the positive association between slaveholding and productivity.

However, correlation does not necessarily imply a causal relationship.[2] More importantly, further examination of the cotton sample suggests that the simple correlation may actually be spurious and that the positive association between slaves and efficiency disappears (and, indeed, becomes negative) after the scale-slaveholding confusion is eliminated. In Table 2 we have controlled for scale in order to reexamine the slave-productivity relationship. The cotton South sample of 4,229 farms was divided into four output classes, and partial and total factor-productivity indexes were recomputed for each of the four F-E slaveholding classes. The table demonstrates that within the output classes each of the productivity indexes *declined* as the size of slaveholding increased. For example, in output class 4 (over $5,000), the total factor-productivity index was 211.3 for slave group three (16–50 slaves). The index for slave group four, however, was only 149.0. For the same slave groups, the output per prime worker index fell from 300.8 to 225.8.

Table 2 suggests that there was no direct relationship between slaveholding and efficiency. It also suggests why this relationship was thought to exist; farms with higher outputs did indeed have higher productivity ratios, and examination of the sample sizes shows that relatively more slave farms were in the high output classes. Hence, what we believe to be a relationship between scale and productivity was misconstrued to be one between slaves and productivity.

To summarize, once we properly distinguished between size of slaveholding and scale, the slaves-productivity relationship proved to be negative. We now turn to a discussion of our proposed link between scale and efficiency.

II. Scale and Efficiency

Fogel and Engerman severely criticized Gavin Wright's contention that their measured slave efficiency correlation was a result of the larger planters' greater emphasis on staple crop production.[3] Preliminary

[2]The linkage between correlation and causation has been widely explored in the literature, for example, see Herbert Simon and Hubert Blalock.

[3]Wright's argument was stated in terms of cotton and corn values only (pp. 316–17). Our measure is preferable because it incorporates other staples and total outputs. Our results are not changed by the use of the simple ratio.

TABLE 2—TOTAL AND PARTIAL PRODUCTIVITY INDEXES BY OUTPUT AND SLAVE CLASSES

Output Class	Slave Class	Sample Size	Partial Productivity Indexes[a] Land	Labor	Capital	Total Factor-Productivity Index[a]
1	1	1638	79.3	69.1	82.4	73.7
	2	434	39.7	52.3	55.2	49.3
	3	12	11.4	15.2	16.9	14.4
	4	0	–	–	–	–
2	1	411	121.2	155.8	128.6	141.6
	2	756	73.6	112.1	86.5	96.6
	3	88	45.7	59.3	60.6	55.8
	4	1	b	b	b	b
3	1	30	248.1	447.2	149.8	320.5
	2	283	103.8	204.5	107.9	154.8
	3	365	87.2	156.0	98.8	124.8
	4	27	38.6	70.8	36.7	54.4
4	1	2	b	b	b	b
	2	11	143.5	563.4	207.8	337.8
	3	138	112.8	300.8	159.8	211.3
	4	103	73.6	225.8	101.7	149.0

Source: Parker-Gallman cotton South sample.
[a]Free Farms = 100.
[b]Calculations for sample sizes less than ten were omitted. See the footnotes to Table 1 for other definitions. Slave classes: 1-no slaves; 2-one to fifteen slaves; 3-sixteen to fifty slaves; 4-more than fifty slaves. Output classes: 1-$1 to $500; 2-$500 to $1,500; 3-$1,500 to $5,000; 4-$5,000 or more.

analysis of the cotton sample was suggestive of a relationship such as Wright proffered—the cash crop to output ratio increased with both output and total factor productivity. Thus, while it is clear from Table 2 that efficiency increased with output, this relationship could conceivably be caused by differences in the crop mix.

Table 3, however, shows that this caveat is unwarranted. We divided the sample into four crop mix classes and computed partial and total factor-productivity indexes for the four output classes used earlier. Large productivity increases were identified within each crop mix class. For example, among farms with between 25 and 50 percent of output in cash crops, total factor productivity (G) increased from 75.2 to 100.9 across the first three output classes. A similar result was found among the staple specializers in the 75 percent or more group. Here, G increased from 82.8 to 187.0 between output classes 1 and 4.

We interpret these findings as strong evidence for economies of scale. Productivity was unmistakably higher for larger farms. Crop mix, however, is not entirely irrelevant. For our data crop mix does not explain productivity differences associated with scale, but it does seem to be related to the *range* over which those differences were realized. For example, among farms whose output was less than 25 percent cash crops, positive productivity differences were exhausted by output class 2. In the next crop mix group productivity rose with output—but at a clearly diminished rate—between output classes 2 and 3. In the remaining two classes there is no evidence that a maximum productivity level is being reached.

Overall, we conclude that F-E were correct in arguing that crop mix could not explain the association between productivity and farm size. They simply erred in gauging "large" in terms of slaveholding rather than output.

III. Partial Correlation Analysis

Our findings can be summarized through correlation analysis. The variables discussed above—output, number of slaves, total factor productivity, and the cash crop

TABLE 3—TOTAL AND PARTIAL PRODUCTIVITY INDEXES BY CROP MIX AND OUTPUT CLASSES

Cash Crop/ Output	Output Class	Sample Size	Partial Productivity Indexes[a] Land	Labor	Capital	Total Factor-Productivity Index[a]
<.25	1	951	43.2	52.8	54.9	50.6
	2	231	49.2	99.4	75.1	79.5
	3	44	47.0	108.4	34.6	72.4
	4	4	b	b	b	b
.25 to .50	1	555	84.8	69.2	84.1	75.2
	2	370	81.5	102.7	81.8	93.3
	3	72	57.3	137.0	81.6	100.9
	4	5	b	b	b	b
.50 to .75	1	464	85.3	73.1	100.3	80.2
	2	498	94.8	119.4	103 9	110.1
	3	362	95.4	155.9	103.0	128.5
	4	86	109.8	210.5	120.0	162.6
≥.75	1	114	94.6	77.7	84.6	82.8
	2	157	92.9	137.1	109.3	119.7
	3	227	103.5	198.2	110.7	152.6
	4	159	95.9	279.6	126.6	187.0

Source: Parker-Gallman cotton South sample.
[a]Free farms = 100.
[b]Calculations for sample sizes less than ten were omitted. Cash crops include cotton, sugar, rice, and tobacco. For other definitions, see footnotes to Tables 1 and 2.

ratio—were calculated for each of the sample farms. The correlation between the number of slaves and productivity r_{sp} was .087.[4] Since our point has been that these estimates ignore scale differences, we computed $r_{sp \cdot o}$ to adjust for output. This statistic has a value of −.259. Hence, the prior positive association between slaveholding and productivity can be "explained away" by scale.

On the other hand, it was not possible to remove the correlation between output and productivity. Letting s = slaves, p = productivity (G), o = output, and c = cash crop ratio, we found

$$r_{op} = .313$$
$$r_{op \cdot c} = .215$$
$$r_{op \cdot sc} = .361$$

The important relationship, therefore, was not between slaves and efficiency, but between scale and efficiency. Adjusting for the crop mix did not remove this correlation, nor did adjusting for the size of slaveholding.[5]

IV. Conclusions

In order to discuss economies of scale it is necessary to define scale in terms of output. Using this preferred definition it was possible to conclude that there were productivity advantages in large-scale production of southern agriculture products. These advantages existed regardless of the crop mix and are not related to the greater staple specialization of large firms. More importantly, we have shown that the F-E interpretation is incorrect. Their observed relationship between slaves and productivity is *explained* by scale of operation.

Our analysis does not tell us why the very largest planters relied almost exclusively upon slave rather than wage labor. A study

[4]All of the reported statistics are significantly different from zero at the .001 level of significance. Also, each of the variables was measured continuously, not discretely in terms of the classes used above.

[5]It should be noted that the same conclusions were reached for the tobacco regions of Kentucky and Tennessee (see Schaefer). Moreover, the conclusions documented here also hold within each of the major soil classifications of the cotton South. For additional evidence of scale economies in other crops, see the studies by Schmitz and Swan.

of the supply of agricultural labor and farm management practices would be required to resolve this issue.[6] Our analysis does make clear, however, that slaves were not the cause of higher productivity. Slaves were *one* means of achieving a larger scale of operation; it was the increase in scale that led to higher levels of productivity.

[6]The most recent comments on this issue are those of Jacob Metzer, Heywood Fleisig, and Ralph Anderson and Robert Gallman. Necessary, but not sufficient conditions for Fleisig's argument that free farms faced an inelastic labor supply would be lower outputs and higher land-labor and capital-labor ratios for free farms than for slave farms. We found that these conditions existed for the sample considered here.

REFERENCES

R. V. Anderson and R. E. Gallman, "Slaves as Fixed Capital: Slave Labor and Southern Economic Development," *J. Amer. Hist.*, June 1977, *64*, 24-46.

Hubert M. Blalock, *Causal Inference in Nonexperimental Research*, Chapel Hill 1964.

H. Fleisig, "Slavery, the Supply of Agricultural Labor, and the Industrialization of the South," *J. Econ. Hist.*, Sept. 1976, *36*, 572-97.

Robert W. Fogel and Stanley L. Engerman, "The Relative Efficiency of Slavery: A Comparison of Northern and Southern Agriculture in 1860," *Explor. Econ. Hist.*, Spring 1971, *8*, 353-67.

——— **and** ———, *Time on the Cross*, Vol. II, Boston 1974.

——— **and** ———, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South," *Amer. Econ. Rev.*, June 1977, *67*, 275-96.

J. D. Foust and D. E. Swan, "Productivity and Profitability of Antebellum Slave Labor: A Micro-Approach," *Agr. Hist.*, Jan. 1970, *44*, 39-62.

J. Metzer, "Rational Management, Modern Business Practices and Economies of Scale in the Ante-Bellum Southern Plantations," *Explor. Econ Hist.*, Apr. 1975, *12*, 123-50.

D. F. Schaefer, "Productivity in the Antebellum South: The Western Tobacco Region," *Res. Econ. Hist.*, forthcoming.

M. D. Schmitz, "Economies of Scale and Farm Size in the Antebellum Sugar Sector," *J. Econ. Hist.*, Dec. 1977, *37*, 959-80.

H. Simon, "Spurious Correlation: A Causal Interpretation," *J. Amer. Statist. Assn.*, Sept. 1954, *49*, 467-79.

Dale E. Swan, *The Structure and Profitability of the Antebellum Rice Industry: 1859*, New York 1975.

G. Wright, "Prosperity, Progress and American Slavery," in Paul A. David et al., eds., *Reckoning with Slavery*, New York 1976.

Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South: Comment

By PAUL A. DAVID AND PETER TEMIN*

The publication of Robert Fogel and Stanley Engerman's (hereafter, F-E) book *Time on the Cross* was greeted by a flurry of laudatory notices in the popular press. Our review article in 1974 began the process of critical reevaluation by economists and quantitative historians which has exposed the serious methodological and substantive errors that permeate *Time on the Cross.*[1] In the original critique, to which F-E (1977) have chosen to reply, we focused attention upon the central position occupied by their "findings" concerning the relative "efficiency" of slave agriculture. The validity of these findings was seen to be of crucial importance within the larger structure of F-E's proposed reinterpretation of American Negro slavery. Based on their conclusion that southern agriculture enjoyed a total factor productivity advantage vis-à-vis free northern farming, they argued that the plantation regime rendered slaves "superior" workers and that blacks today should take pride in the accomplishments of their captive forebears whose diligence and skill made American slavery an economically efficient system. Putting aside the difficult factual, ethical, and emotional questions surrounding the extent of acquiescence (or active cooperation) in the plantation system by slaves, we argued that F-E's method of demonstrating (southern) plantation agriculture's "relative efficiency," compared with free (northern) farming, was badly flawed. Indeed, we found conceptual and empirical errors so grave that they invalidated F-E's findings on this key issue.

Fogel and Engerman (1977) have returned to their productivity calculations in an attempt to reconstruct this foundation of their argument, but their attempt only further undermines their intellectual edifice. They did not discuss the theoretical objections to their calculation at all, and they performed new calculations that are even more distorted than the old. In responding to our criticisms of the way they measured land inputs in calculating relative factor-productivity averages for farms in the North and South, and for different classes of farms and plantations within the South, they have entirely eliminated land as an input in the agricultural production function. Although the results of these productivity recalculations have been seized upon by F-E to argue that our theoretical objections are of no empirical importance, these new "corrected" figures compound rather than correct the original error.

We substantiate these propositions in three steps. First, we restate and clarify the theoretical objections to F-E's procedures. Second, we demonstrate that F-E (perhaps inadvertently) have removed land from their agricultural production function. And third, we show the effect of a rudimentary correction of their productivity figures that at least moves in the right direction.

The issue of efficiency is important to F-E because they based their argument that slaves were more productive workers than free white farmers on the presumed greater efficiency of southern agriculture. As employed in this argument, greater efficiency

*Professors of economics, Standord University and Massachusetts Institute of Technology, respectively. We would like to thank Gavin Wright for comments. David would like to thank the National Science Foundation and Temin would like to thank the Charles Warren Center for Studies in American History, Harvard University, for financial support.

[1] Fogel and Engerman's (1977) reply to their critics overlooks the more comprehensive appraisal of *Time on the Cross*, published under the title *Reckoning with Slavery*, in which we were joined by Herbert Gutman, Richard Sutch, and Gavin Wright.

refers to superior technical efficiency, the ability to produce a specified set of outputs using a smaller amount of inputs than that required by another productive entity. But F-E measured the efficiency with which southern and northern farms generated *revenue*—what we call "revenue efficiency"—which differs conceptually from technical efficiency. Relative technical efficiency can be inferred from relative revenue efficiency under certain circumstances, but not under those which are appropriate to the historical case at hand.

The output mix of southern and northern farms, and even of different sized farms within the South, was very different. In particular, no cotton was grown in the North, and smaller southern farms had smaller proportions of cotton in their output. In their measurement of efficiency, F-E aggregated different farm products in a fixed-weight index, with national market prices as the weights. Their implicit assumption was that only the size of this aggregate index, and not its composition, mattered in establishing relative levels of technical efficiency for slave and free farms, and for southern and northern agriculture. This would be true if the relevant portion of the production-possibility frontier describing combinations of farm output attainable from a given set of inputs were linear and parallel to the slope of relative prices, but it is not true if the production-possibility frontier was curved.

Consider Figure 1. The two axes measure the production of cotton and "corn," the latter being taken here as a proxy for all noncotton products. The line *AB* is drawn to reflect the ratio of (assumed) market prices. If it also indicates the shape of a representative production-possibility frontier (using a standard input bundle), then the relative locations of the production-possibility frontiers for different types of farms are shown unambiguously by which price lines their respective outputs lie on. Output on a higher price line indicates that production was on a higher production-possibility frontier.

On the other hand, if the production-

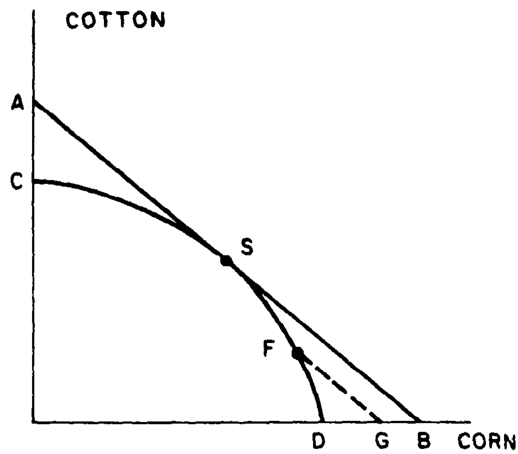


FIGURE 1

possibility frontier was curved, as shown by *CD* in Figure 1, this correspondence between the relative positions of price lines and production-possibility frontiers does not necessarily exist. It exists only if the different producers faced the same relative prices *and* were profit maximizers, for then we can assume that each one produced at a point analogous to point *S* in Figure 1. (*S* is the point of maximum revenue, and therefore, given the cost of the standard input bundle, the point of maximum profit). But if different producers faced different prices, or if they were prevented by climactic reasons from producing the same composition of output, or if they took risk or some other thing extraneous to the analysis of Figure 1 into account in determining the composition of output, then the simple correspondence between price lines and production-possibility frontiers is absent. By definition, every point on the production-possibility frontier *CD* is technically efficient: *F* and *S* show points of equal *technical* efficiency. Yet *F* is less *revenue* efficient than *S* at the prices indicated by *AB*.

In the case of antebellum agriculture being discussed here, the assumptions needed to infer the existence of greater or lesser technical efficiency from observations of revenue efficiency clearly are absent. Cotton could not be grown in the North; farms

in that region were restricted to the X axis of Figure 1. Further, a succession of investigators has demonstrated that the supply of cotton was almost completely price inelastic; since the South did not produce cotton exclusively, the southern production-possibility frontier could not have been linear.[2] Therefore, equal *technical* efficiency in "corn" production in the North and South would imply higher *revenue* efficiency in the cotton producing South.

Within the South, small farmers either faced different prices or took into consideration factors other than those considered in plantation managers' decisions, since the proportion of different crops grown differed among farms in the South.[3] The measure of productivity used by F-E shows the ability of different types of farming units to generate revenue at national market prices. It says nothing about the relative location of production-possibility frontiers, and a fortiori, nothing about the technical superiority or inferiority of particular factors of production such as slaves and free farm laborers.

On the contrary, as we and others have argued (see David et al., chs. 5, 7) and as we will reconfirm below, measured differences in revenue efficiency arise predominantly from differences in the ability or willingness to grow cotton, not from any presumed difference in the efficiency of slave and free labor.

The argument can be clarified by reference to Figure 1. Let *CD* in Figure 1 represent the southern production-possibility frontier. Then *S* shows where profit-maximizing plantations produced, and *F* indicates the position of equally efficient small southern farms who produced relatively less cotton than the large plantations. Northern farms producing on the corn axis (since cotton would not be grown in the North) at *D* would be less revenue efficient than southern farms at *F* or *S*, but no less technically efficient. Since we do not know the exact shape of the southern production-possibility frontier, we cannot infer from the location of a point on the corn axis whether it lies above or below the southern production-possibility frontier—unless it lies at or beyond a point like *G*. Given the uniform price ratio between cotton and corn indicated by *AB*, a northern farm producing at *G* would appear on a revenue efficiency par with a small slave farm producing at *F*, and less revenue efficient than a large plantation more specialized in cotton production at *S*. Yet, given a southern production-possibility curve passing through both *S* and *F*, the northern farm at *G* was actually technically superior to the southern farms in the only activity in which both regions could engage—corn production.

Fogel and Engerman argued in their recent article (1977) that our insistence on the importance of cotton is misplaced and that their calculations show this theoretical problem to be insignificant. Starting from our contention that using market values of land as a measure of land inputs underestimated the land input of all but the largest and best located southern plantations in their original productivity calculations, they made an allowance for the depressing effect on southern land values of the poor southern transportation system, reestimated the southern land input, and recalculated agricultural productivity for different sized southern farms. Since their new results showed that farms of two different sizes (having different product mixes) had the same productivity, they concluded that the curvature problem just described was unimportant.

We now turn to their calculations and their new estimates of relative productivities. We will demonstrate that F-E's new figures are even more biased than the old ones and that an appropriate adjustment

[2]See Temin, Wright, and Stephen DeCanio. Even if these investigators have underestimated the elasticity of supply of cotton, it would be hard to argue that the Southern transformation curve was flat. See Jacob Metzer for a demonstration that the labor requirements of cotton and corn were complementary.

[3]Alternatively, the production-possibility frontier might have had a different shape for different sized southern farms. If so, then the argument about differences within the South is exactly like the argument on North-South differentials.

for the lack of transportation facilities in the South (using F-E's own data) confirms the importance of our criticism.

To understand Fogel and Engerman's calculations, it is necessary to start with their equation (1), which defines their efficiency index for the South. We reproduce it here for convenience:

$$(1) \quad G_s/G_n = \frac{Q_s/Q_n}{(L_s/L_n)^{\alpha L}(K_s/K_n)^{\alpha K}(T_s/T_n)^{\alpha T}}$$

This equation can be rewritten as follows, letting X_i stand for all the terms that do not involve land (T) and altering the subscripts to reflect F-E's concern in their article with productivity differences among different sized farms *within the South*:

$$(2) \quad G_i/G_o = \frac{X_i}{(T_i/T_o)^{\alpha T}}$$

Fogel and Engerman (1977) did not give their correction formula explicitly, but it can be inferred from examination of the data in their Tables 6 and 7. The first column of Table 7 gives the relative productivity indexes as they appeared in F-E (1974). The second gives a "corrected" series. If we label the entries in Table 6, a_{ij}, and the entries in Table 7, b_{ij}, the calculation involved in getting from b_{i1} to b_{i2} can be represented as follows:

$$b_{i2} = (a_{91}/a_{9i})^{1/4} b_{i1}$$

where a_{91} and a_{9i} are the appropriate entries in row (9) of Table 6. This formula can be verified by performing the appropriate calculations.

The meaning of this correction factor can be seen by expressing it in terms of the symbols of equations (1) and (2). The fourth root was used because α_T was set equal to .25 by F-E when they employed equations (1) and (2). The ninth row of Table 6 gives F-E's new estimates of the ratio of the value of improvements to the total value of land and improvements, V_i/T_i. The denominator (T_i), of course, was F-E's original measure of land input. The correction formula, then, is

$$(3) \quad k_i = \left(\frac{V_o/T_o}{V_i/T_i}\right)^{\alpha T}$$

where V_i is defined to be the value of improvements for farm size i.

Multiplying G_i/G_o as defined in equation (2) by k_i reveals the effect of F-E's correction:

$$(4) \quad \frac{G_i}{G_o} k_i = \frac{X_i}{(T_i/T_o)^{\alpha T}} \left(\frac{V_o/T_o}{V_i/T_i}\right)^{\alpha T} = \frac{X_i}{(V_i/V_o)^{\alpha T}}$$

Fogel and Engerman's correction of their original productivity estimates consists of simply removing the value of land and improvements and replacing it by the value of improvements alone. In response to our criticism of their valuation of land, F-E have entirely eliminated land inputs from their production function!

This clearly is inappropriate from a theoretical point of view. An agricultural production function that does not include land is exceedingly hard to interpret.

From the practical point of view, F-E have responded to our criticism that their measures of the land and improvements input for all but the largest plantations (and therefore for southern agriculture as a whole) were too low, by *reducing* them still further. The new calculations therefore move the relative productivity indexes for slave farms in the South upwards, in just the opposite direction to that called for by criticism which Fogel and Engerman appear to accept as justified.

It is not our purpose here to redo F-E's productivity calculations properly. Indeed we argued originally and have been obliged to reassert here that the conceptual apparatus they are using is inappropriate to the task of ascertaining the relative technical efficiency of free and slave agriculture. Nevertheless, we cannot resist exploring the implications of a correction that would at least move the productivity calculations in the right direction. Fogel and Engerman (1977) argued that the difference between the value of unimproved southern land in

the largest plantations and in all other size farms (shown in row (4) of their Table 6) was due to location. It follows that the value of well-sited plantation land should be used instead of the actual value of land in smaller, less well-located farms to correct for the locational disadvantage of these smaller southern farms.

In other words, the value of land used in the original calculation should be replaced by the product of the average number of acres in farms and the price of unimproved land on well-located plantations. Since F-E's figures show the price difference between well-located and their actual unimproved land to have been approximately \$10 an acre for all southern farms other than plantations, we can get corrected values of T_i by adding ten times the average number of acres in a farm to the actual value.

Substituting the corrected values of T_i for F-E's in equation (2) yields the estimates shown in the first column of our Table 1, which may be compared with the "corrected" indexes from F-E's Table 7. The productivity of the smallest southern farms, those without slaves, is defined as 100 in both cases.

The previous identity between the productivity levels calculated for free farms and the small (1–15) slave farms, which F-E relied upon to refute the argument that the apparent productivity advantage of southern slave farms was due in large part to their relative concentration upon growing cotton, has disappeared. Indeed, the corresponding revision of the ratio of the efficiency index for cotton (A_c) to the efficiency index for other crops (A_o)—calculated by F-E's own methods, using the crop mix data for farms in these two size categories—now is found to be 3.1 instead of the ratio of 1.1 that F-E presented. The range of alternative estimates of this ratio (A_c/A_o) now runs from 3.1 to 10.8. The variance of these numbers is not surprising, given the roughness of our correction, and each separate calculation confirms the hypothesis that cotton was more efficient than other crops (recalling always the unusual meaning of efficient in this context).[4]

TABLE 1—REVISED "EFFICIENCY" FIGURES FOR SOUTHERN FARMS (Free Farms = 100)

	Corrected for Omission of Land	F-E (1977)
Free farms	100	100
1-15 slaves	114	101
16-50 slaves	159	133
51+ slaves	174	148

In addition, the calculated productivity of free southern farms is reduced—not raised as F-E's (1977) calculations would have us believe—by this adjustment. It falls to four-fifths of its old level [$(1/2.92)^{1/4} = .77$)]. Free southern farms now appear to be *less* revenue efficient than free nothern farms, despite the advantage they had in being able to grow cotton—given the prevailing market prices of that crop. Even though the productivity position of the large plantations changes relative to other southern farms, their position relative to that of northern farms is unaltered by virtue of the adjustment we have carried out. This is so because the value of the large plantations' lands was assumed not to have been depressed by poor transport facilities. It follows, therefore, that some southern farms were more revenue efficient than northern farms, and some less. Were the interregional comparisons to be made farm by farm, rather than by preaggregating the input and output data, the variance among the productivity ratios for individual farming units undoubtedly would be so large as to preclude finding any statistically significant differences in revenue efficiency between the major regions.

On the premise that the southern production-possibility curve had the general curvature indicated in Figure 1, the revenue efficiency comparison which places northern farms at a point between G and B in

[4]The estimates for A_c range from 2.54 to 4.86, with mean 3.63. The estimates for A_o range from .45 to .88, with mean .62. In no case is the estimate of A_c close to the estimate of A_o, as the range of estimates confirms. Far from refuting our contention, the revised data—rough as they are—confirm it most strikingly.

Figure 1 permits us to infer that northern farms were more *technically* efficient than southern farms in the production of corn. Indeed, the calculated inferiority of free southern farms in terms of revenue efficiency must understate the extent of their technical efficiency disadvantage vis-à-vis northern farms, since the former enjoyed the revenue benefits of being able to crop some cotton.

We do not present the results of our revised productivity calculations with any claims to finality. Comparable adjustments of the land input measures for the North could be made, and other changes as well.[5] The revisions we have undertaken simply rectify the most recent distortions introduced by F-E and reestablish the importance of crop mix effects upon measures of relative revenue efficiency.

Thus, the concluding point on which we want to insist is the difference between the concepts of revenue efficiency and technical efficiency when the production-possibility frontier is curved. Those southern farms that are found to be more revenue efficient than northern farms were benefiting from their ability to grow cotton. It is not possible to say precisely how much of the calculated "efficiency" differences in such cases is due to growing cotton, rather than to the managerial skills and methods of organization adopted by planters and overseers, or to the sheer hard work extracted from slave labor. Our contention is that it was due almost entirely to cotton. That interpretation is supported not only by the econometric evidence on crop mix effects previously adduced by Wright (see David et al., ch. 7), but has been seen here to be entirely consistent with the data presented in F-E's latest contribution.

[5]For example, considering just the matter of land inputs, our calculations imply that unimproved southern land was worth approximately six times as much as northern unimproved land, once the locational disadvantage is eliminated. This might well be a major overstatement. On the other hand, it could also reflect the realities of 1860. The North was well settled, and only poor or inaccessible land in the North was still unimproved.

REFERENCES

Paul A. David and P. Temin, "Slavery: The Progressive Institution?," *J. Econ. Hist.*, Sept. 1974, *34*, 729–83.

——— **et al.,** *Reckoning with Slavery*, New York 1976.

S. DeCanio, "Cotton 'Overproduction' in Late Nineteenth-Century Southern Agriculture," *J. Econ. Hist.*, Sept. 1973, *33*, 608–33.

Robert W. Fogel and Stanley L. Engerman, *Time on the Cross*, Boston 1974.

——— **and** ———, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South," *Amer. Econ. Rev.*, June 1977, *67*, 275–96.

J. Metzer, "Rational Management, Modern Business Practices, and Economies of Scale in the Ante-Bellum Southern Plantations," *Explor. Econ. Hist.*, Apr. 1975, *12*, 123–50.

P. Temin, "The Causes of Cotton-Price Fluctuations in the 1830's," *Rev. Econ. Statist.*, Nov. 1967, *49*, 463–70.

G. Wright, "An Econometric Study of Cotton Production and Trade, 1830–1860," *Rev. Econ. Statist.*, May 1971, *53*, 111–20.

The Efficiency of Slavery: Another Interpretation

By GAVIN WRIGHT*

The managing editor of this *Review* has very kindly invited me to comment on the recent reiteration by Robert Fogel and Stanley Engerman (hereafter, F-E) of their views on the efficiency of American slave agriculture (1977). Based on my reading of the article, my conclusions are that their lines of defense do not adequately respond to the most important objections, and that their new interpretations suffer from many of the shortcomings of the original work. This brief note will discuss two problems: the representativeness of the census year 1860, and the effects of the crop mix on the apparent efficiency of slavery. A concluding section will suggest some alternative lines of analysis for future work.

I. The Census Year 1860 in Perspective

In *Time on the Cross*, F-E argued that southern agriculture was more efficient than northern agriculture in 1860, and that this regional superiority resulted from an efficiency advantage of slave-using enterprises over free farms in both regions. They interpreted this efficiency advantage as a superiority in physical efficiency, which they attributed to economies of scale in plantation operations and to the development of distinctive managerial skills and methods among planters. The same factors were held to be responsible for the enormous expansion of cotton production and the rapid growth of southern per capita income down to 1860. The efficiency calculations on which these conclusions were based (and the subsequent revised estimates presented in their 1977 article) used data from only one crop year, the census year 1859-60. In response to this analysis, along with several other writers, I argued (presenting econometric evidence) that the cotton demand situation was unusually favorable during 1820-60, especially for the census year 1860. I also argued that the cotton yield of 1860 was unusually high, because output was far above the level predicted by a supply curve estimated over the period 1820-60.

In their rebuttal to the critics, F-E reject these results (1977, pp. 280-82). They treat the two arguments as though they were entirely separable, dismissing the importance of demand solely by reference to the relative cotton price, and disposing of the supply argument with a discussion of yield variability. But this is an elementary error: because the world cotton price was heavily influenced by the size of the American crop, price is not a measure of demand.[1] And because *both* supply and demand were abnormally high, the price was not exceptional, but the year in question was exceptional to say the least.

To show how evident these points are, and to indicate that they do not depend on some special econometric formulation, I display as Figure 1 F-E's own index of cotton supply and demand, reproduced from page 92 of *Time on the Cross*. The unique position of the census year 1860 for both supply and demand is apparent. Fogel and Engerman dispute the assertion that per acre cotton yields were unusually high in 1860, but their discussion is largely irrelevant: they offer no alternative reason why one particular year should be so far above the trend, and the precise reason for

*University of Michigan.

[1]The elasticity of demand for American cotton has been estimated at approximately unity, if not lower. Because the cotton crop was essentially predetermined, the demand curve may be estimated by ordinary least squares with price as the dependent variable. The simplest estimate for the antebellum period is

$$\ln P = 8.21 - 0.94 \ln Q + 0.052$$

which implies an elasticity of demand equal to 1.06. The residual from this curve in the census year 1859-60 is 15.9 percent.

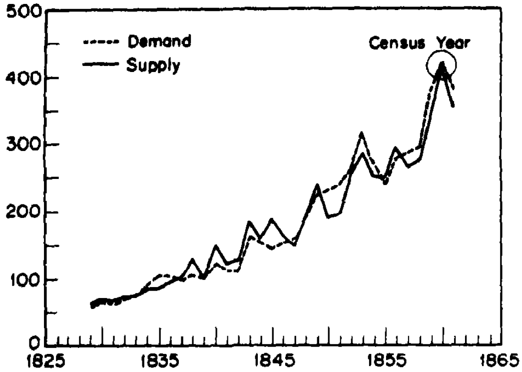


FIGURE 1. A COMPARISON BETWEEN INDEXES OF COTTON DEMANDED AND SUPPLIED, 1829 61

this upward deviation is immaterial. If yields were not unusually high, the alternative interpretation can only be that an unusually large cotton acreage was planted in that year; but such a development, in conjunction with favorable demand, in no way challenges the assertion that the crop year in question was atypical. Fogel and Engerman stress that relative cotton prices were even higher in 1850 than in 1860, but they do not mention the reason, which was that the 1850 crop was a particularly bad one which "encountered a series of disasters from first to last."[2] The claim that more than 91 percent of the increase in cotton production between 1850 and 1860 can be "accounted for" by long-run factors (1977, p. 281) is rather implausible, as readers may see by drawing a line connecting the two census year outputs and comparing its slope with that of a trend line for the whole period.

The significance of the extraordinary character of the census year 1860 is not mainly for the calculation of aggregate regional indices of efficiency.[3] But because the relative importance of cotton was so much greater for large slave plantations than for small southern farms, one would expect an unusually favorable cotton year to affect cross-sectional comparisons within the South. Fogel and Engerman's figures confirm that the average share of cotton in the gross value of farm output varied from 29 percent on slaveless farms to 61 percent on plantations with more than 50 slaves (1977, p. 288). Even if the high cotton output of 1860 were attributable to a large cotton acreage rather than yields, the cross-sectional pattern would still be affected: the precise effect would depend on the distribution of "abnormal" cotton acreage among farm sizes, but since roughly half of the slaveless farms grew no cotton at all, they obviously could not have shared in the distinctive prosperity of that year. And since available evidence indicates that the marginal return to cotton acreage (measured at market prices) was substantially above that of corn acreage in all but the poorest years, the *combination* of heavy cotton planting and strong demand will markedly effect an efficiency calculation which ignores these facts.

Despite its secondary importance, however, it is worth looking more closely at F-E's denial that yields were unusually favorable in 1860. This denial is based on their belief that my estimate of the upward deviation in yields is too high to be believable (1977, p. 281). On its face this is a rather peculiar argument: they assert that the residual for 1860 is so large as to occur only once in a thousand cases, they argue that this is so unacceptably high as to completely invalidate the underlying supply curve estimate, and they apparently conclude that they can safely proceed as though 1860 were just an ordinary year. But this is also an inaccurate argument, because they compare the 1860 residual for a five-state area (in which the high yield was concen-

[2]See U.S. Commissioner of Patents (1849, p. 149). Contemporary accounts of the crop failure are abundant: see Commissioner of Patents (1849, Vol. 2, pp. 144, 170, 307); Commissioner of Patents (1850, Vol. 2, p. 510); James Watkins (pp. 81, 197, 150, 197, 217, 240, 258). Note that the cotton crop for the census year 1850 actually grew in the calendar year 1849.

[3]I omit discussion of these regional calculations because in my opinion the conceptual problems which afflict such comparisons are insuperable. Since in any case the South's superiority is entirely attributed by F-E to the superiority of slave plantations over the free farms of both regions, the entire argument rises or falls with the cross-sectional comparison within the South.

trated) with the 1867–1900 standard deviation for the entire South;[4] my residual estimates for the whole South in 1860 (11.6 percent to 23.6 percent) could hardly be called incredible.[5] It is, finally, an inappropriate argument because the probability calculations assume that yields are normally distributed, when in fact the distribution of cotton yields is known to be positively skewed (see Richard Day).

It is in the nature of cotton yield determination that this should be so. Fogel and Engerman appear to believe that the yields of all crops are highly correlated, and they implicitly scoff at the notion that the "favorable weather that supposedly visited the South in 1859–60" (1977, p. 282) should have affected only cotton; but the rainfall and temperature requirements of cotton are very distinctive and very demanding, and cotton yields are not highly correlated with those of other crops even in the South.[6] Fogel and Engerman's estimates of the "reallocation of corn land to cotton" (1977, p. 282) between 1850 and 1860 is based on the implicit assumption that relative outputs of the two crops were proportional to relative acreage planted in the two years (i.e., that yield fluctuations were identical for cotton and for corn). These estimates are highly misleading, because despite the abundant testimony to the failure of the 1850 cotton crop, the corn crop for that year was quite a good one, and observers noted: "Cotton was hurt by the same rains that helped corn" (p. 64).[7] Cotton passes through several critical phases of development during the growing season, and bad weather at any one of these points can drastically reduce yields below potential. Even if the growth of the stalk and bolls are ideal, it was common in the nineteenth century for the fruit to deteriorate rapidly from inclement weather during the picking season.[8] Periodically, however, a favorable growth capped by an extended period of fair weather for the harvest will produce a

[4]The postbellum figures show that the standard deviation for the five-state area is 20 percent higher than the aggregate standard deviation.

[5]Fogel and Engerman's objections to the supply curve itself (1977, p. 281) are also inappropriate. The most general specification is

$$\ln Q = A + b \ln P_{-1} + ct + d \ln L_{-2}$$

where L_{-2} is the cumulative acreage of public land sold in the cotton states, lagged two years. This equation does not assume that, for example, the "proportion of land in farms that was improved" was "fixed during the decade of the 1850s" (1977, p. 281), nor does it imply such constancy about any of the ratios which they list. None of these sources of output growth were new to the 1850's, and they are all presumably reflected in the coefficients of P_{-1}, L_{-2}, and certainly t. I would be the last to claim that an equation like this one embodies any very deep explanation for the growth of cotton output. The whole point of including land sales is simply to account for more of the year-to-year variation in the growth of cotton acreage and hence to obtain a better estimate of yield fluctuations. Contrary to what F-E imply, these equations track the growth of output very well during the 1850's—except for the census year deviation. Econometrically speaking, the main effect of including lagged land sales is to eliminate the pattern of serial correlation in the residuals which is obtained when the *log* of output is regressed against time. If the land sales variable is considered objectionable, the same effect (and similar residual estimates for 1850 and 1860) can be generated by conventional methods of correction for autocorrelation. Examination of residual plots from any of these regressions (available from the author on request) leaves no doubt as to the distinctiveness of the census year under discussion.

[6]Fogel and Engerman compute a peculiar parameter, "the elasticity of corn yields with respect to cotton yields" (1977, p. 281, fn. 12), but using the same data one finds that only one-quarter of the variation in cotton yields is explained by variation in corn yields. Much better data from twentieth-century experiment station records show no correlation whatsoever (r^2 = .0036). The data may be found in Perrin H. Grissom and W. I. Spurgeon.

[7]See Alfred Smith, pp. 63–65. Referring to South Carolina, Smith writes that the heavy rains "helped the corn considerably, and the crop was a good one, one of the best ever grown" (p. 64). Comparison of the 1850 census crop with the Patent Office estimates for the preceding year shows that cotton output declined while the corn crop (in the seven leading cotton states) rose by 5 to 10 percent.

[8]"Normally the cotton plant produces bolls the entire length of the stalk.... It often happens, however, that unfavorable weather conditions cause a part of the fruitage to fall, and if bad weather prevails for a sufficient period a considerable portion of the stalk may become bare of fruit.... In fact, it seldom happens that a season is so favorable that the plant is fruited from bottom to top, but when such is the case a bumper crop is the result" (see James Covert, p. 93).

phenomenal yield, far above average. This seems to have been what occurred in the Southwest in 1859–60.

Fogel and Engerman assert that "available commentaries" on the 1860 crop "are devoid of references to an extraordinarily high yield" (1977, p. 281), but numerous testimonials are in fact available. Ezekiel Donnell's survey of the New Orleans press, for example, cites the "fine and very favorable picking weather" as the explanation for the cotton receipts "larger than ever before known" in December 1859 (p. 496). The *American Cotton Planter* of Montgomery states that "the picking season has been one of the most favorable ever known, especially in the Southern states" (p. 463). Even the rainfall data attest to the unusually fine weather during the harvest season.[9] And after noting that the fine weather had probably improved the quality of the cotton crop as well, the writer for the *American Cotton Planter* went on to marvel that "the prices have been wonderfully sustained." He concluded: "Taken as a whole, the cotton interest was never in a more prosperous condition" (p. 164). To avoid misunderstanding, there is no claim here that F-E's results would be reversed by data from another crop year. But the choice of year clearly exaggerates the apparent efficiency of slave labor, and it is unreasonable to deny or ignore this fact.

II. The Crop Mix Effect

In my earlier critique, I argued and presented evidence that F-E's efficiency index was closely linked to cotton, and that the share of cotton in output was the main determinant of the value of output per worker (1976, pp. 316–18; 334–36). I suggested that this pattern of crop selection could be explained by the riskiness for small farms of growing cotton at the expense of food crops for home consumption. In their response, F-E reject this interpretation on the grounds that the correlation between cotton and efficiency, using their revised calculations, is less than perfect (1977, p. 289). They go on to express skepticism about the proposition that cotton was a riskier choice than corn, and to argue that the figures imply an implausibly large risk premium in exchange for "some unspecified reduction in the variance of ... income" (1977, p. 290). On this basis they apparently conclude that no account whatever need be taken of the effects of crop mix on efficiency.

Nothing in this discussion would lead me to alter my views. It is true that the new efficiency figures would somewhat reduce the correlation between cotton and efficiency; but these new figures have been incorrectly calculated, as Paul David and Peter Temin show. The key new result is that free farms and small slave farms are now found to have been of equal efficiency when locational rents are removed; but taken at face value, the new estimates do not indicate that small slave farms had any locational advantage over free farms. Their estimates place the average value of unimproved acreage, which they identify as "the locational component of land" (1977, p. 283) at $2.57 per acre for slaveless farms, $2.53 for farms with one to fifteen slaves (1977, p. 284, Table 6, line (4)). Hence, these figures do not suggest that any alteration in the relative efficiency ranking of these two classes is called for. But even if this were not the case, it is difficult to see why we should limit outselves to four size-class aggregates when the micro data them selves are available, and why a failure to find universal constancy in efficiency should call for complete dismissal of the crop mix effect. Direct statistical tests show the importance of the share of cotton in total production as a determinant of productivity when output is valued at market prices. This effect is revealed in the following semilog regression run on the Parker-Gallman sample of farms in the cotton South in 1860:

[9]During the harvest months October through January, 1859-60, rainfall in Mobile was only half of the average for the nineteenth century for these months; only three other years fell below this figure (in the period for which records exist 1840–1900). For New Orleans, rainfall was less than two-thirds of the nineteenth-century average. The data may be found in H. Helm Clayton.

$$(1)\quad ln(V/L) = 4.87 - \underset{(0.4)}{.0004SQ} + \underset{(26.45)}{1.27^{**}CS} - \underset{(1.78)}{.0001IA}$$
$$R^2 = .129; N = 4977$$

where V = value of crop outputs, L = labor, SQ = index of soil quality, CS = share of cotton in total crop output, IA = improved acreage, and t-ratios appear in parenthesis. Experimentation with additional explanatory variables has not uncovered any which eliminate or even seriously weaken the effect of CS on $ln(V/L)$.[10] To be sure, F-E subject their input-output figures to a battery of elaborations and refinements, but nothing in their discussion challenges this basic relationship.

Indeed, F-E seem to recognize such a relationship, but they acknowledge only that the "optimum share of cotton" (p. 289, fn. 26) will differ for each slaveholding size class. What they do not mention is that the same relationship holds when the sample (for the same year and geographical area) is restricted to *slaveless southern farms alone:*

$$(2)\quad ln(V/L) = 5.01 + \underset{(17.97)}{1.27^{**}CS} + \underset{(4.17)}{.010^{**}SQ}$$
$$R^2 = .123; N = 2412$$

Note that the coefficient for the cotton share is precisely the same as in the aggregate regression (1). Thus, even if it were true that the "optimum cotton share" were higher on large plantations than on small farms, the fact remains that the crop mix per se had a marked impact on productivity at all farm sizes, as one would certainly expect in a bumper cotton year. The significance of this effect does not depend on acceptance of any particular formulation of the risk argument: it is not particularly critical whether equations (1) and (2) reflect risk preference, leisure preference, irrationality, or merely the luck of the draw in 1860. The important point is that the crop mix called the tune in that year.[11]

Since the question of risk has been raised, however, F-E's discussion deserves a brief response. It is a disappointment to read the complaint that my meaning is "never clearly defined" (1977, p. 290), when they fail to cite the background article to which readers were referred for explanation (my paper written with Howard Kunreuther),[12] and they do not acknowledge the definition and empirical evidence presented there.[13] In that article it is made clear that we are not discussing the "variability of income," but the risk of falling below self-sufficiency in

[10]For example, adding IA/L to the regression significantly improves the fit but has no effect on the coefficient of CS:

$$ln(V/L) = 3.02 + \underset{(6.33)}{.0057^{**}SQ} + \underset{(30.88)}{1.28^{**}CS} + \underset{(39.31)}{.606^{**}IA/L} - \underset{(9.88)}{.005IA}$$
$$R^2 = .335; N = 4969$$

In fact, we now have a significantly negative relationship between productivity and scale.

[11]The results also show the fallacy in F-E's argument that the regional crop mix problem can be waved away because small southern farms, which had the choice of growing cotton, are found to have efficiency levels equal to the average of northern farms (1977, p. 288). That argument assumes explicitly that small southern farms chose an efficiency-maximizing crop mix in 1860. ("But since there was no climatic obstacle that prevented the free southern farms of the cotton belt from choosing exactly the same mix of products that was selected by large slave plantations, it may be assumed that they chose the product mix that was most efficient for them. Presumably a product mix with a larger cotton share would have decreased, or at least not increased, their efficiency" (1977, p. 288). But regression (2) suggests that small farms would have markedly increased their productivity if they had planted more cotton, even without acquiring slaves.

[12]The referrals appear in my papers: 1975, p. 446; 1976, p. 318.

[13]It is similarly disappointing that they raise the possibility of a cash or liquidity constraint and state: "Wright does not explore this issue, although it is widely suggested in the traditional literature on nineteenth century agriculture" (1977, p. 290, fn. 30). But this is one of the central themes of our uncited article. The notion of a cash constraint is widely suggested in the literature on *post-Civil War* southern agriculture and may certainly help to explain the concentration on cotton by small farmers after 1865. It is not of much help in explaining the very limited amounts of cotton grown by small farmers in 1860, though the effort to avoid being caught in a treadmill of indebtedness is one of the underpinnings of the safety-first model.

basic foodstuffs.[14] We argued that because certain food requirements must be satisfied regardless of a particular year's yields and prices, relevant measures of risk (of this type) should be denominated in units of food, primarily corn in the historical case now under discussion. In this context, we compared two alternative means of obtaining corn: production of corn for on-farm consumption, or production of cotton in exchange for purchased corn. Assuming that corn required no cash inputs, we defined the two choices as (in corn-bushel-equivalents per acre):

$$x = \frac{Y_{cot}P_{cot} - D}{P_{crn}}$$

$$y = Y_{crn}$$

where Y_{cot} = cotton yield, Y_{crn} = corn yield, P_{cot} = cotton price, P_{crn} = corn price, D = additional costs per acre associated with growing cotton instead of corn. Using postbellum data Kunreuther and I showed that the standard deviation of x was four to five times greater than the standard deviation of y (see p. 537). We also showed (pp. 534-535) that crop allocation choices in 1860 could be explained by farm size and by subsistence requirements, as measured by number of family members per improved acre. In fact, farm population per improved acre was twice as great on slaveless farms as on the largest class of slave plantations.[15] Clearly the small farms had to plant a larger relative acreage in food crops, even if slave plantations also sought self-sufficiency at the same per capita consumption levels. And as the safety-first model would predict, per capita corn production remained virtually constant between 1840 and 1860, as the expansion of acreage beyond food requirements went predominantly into cotton.[16]

Thus, when F-E document the high variability of the corn price compared to the cotton price (1977, p. 290 and fn. 28), they are supporting my position and not theirs. The many uncertainties of *buying* corn were precisely the reasons why farmers chose to grow their own. And, when F-E claim that the implied risk premium was implausibly high, they forget completely about the exceptional character of the year in question. The census year 1860 was outstanding for cotton but was reported to have been only poor to medium for corn.[17] In contrast, as previously shown, the census year 1850 was characterized by "excellent crop except cotton."[18] A rerun of regression (1) for a sample of cotton South farms of all sizes from the 1850 Census shows the following result:

$$(3)\quad ln(V/L) = 4.92 + \underset{(1.84)}{.009SQ} + \underset{(4.08)}{0.50^{**}CS} + \underset{(0.89)}{.000081A}$$

$$R^2 = .050; N = 668$$

The crop mix effect remains significant, but the coefficient value is less than 40 percent of its 1860 level. Whereas in 1860 the favorable crop mix effect is found in all regions and size classes, in 1850 these subset regression coefficients are frequently insignificant and are sometimes negative. As Kunreuther and I showed for the postbellum years, while cotton was a more profitable choice on the average, there were a number of years for which farmers would have saved more money growing corn than they made

[14]This target has little to do with modern estimates of nutritional requirements (see F-E, 1977, p. 290), but refers instead to the farm household's own assessment of its minimum tolerable standards. Note that because our interpretation hinges on the essentially financial consequences of shortfalls (indebtedness or loss of assets), we are including the feed requirements of working livestock in the determination of minimum requirements.

[15]For the cotton South as a whole, farm population per improved acre was .225, .153, .133, and .119 for the four slaveholding categories, respectively (0, 1-15, 16-50, over 50).

[16]For the five cotton states of the deep South, corn output per capita stood at 29.07, 31.07, and 29.58 in the three census years.

[17]See the citations in Robert Gallman (1970, p. 8).

[18]See W. I. Thorp, p. 125. The Patent Office crop estimates for the 1840's show a marked *rise* in the corn crop between 1848-49 and 1849-50, while the cotton output figures show a decline. See Gallman (1963)

growing cotton (see p. 537). Thus if the risk premium for 1860 appears to be implausibly large, the explanation is simply that the farmers did not know in advance that the year would turn out that way.

III. Some Alternative Lines of Interpretation

Taking a long view of the matter, the special features of the year 1860 are less important than the special features of the whole antebellum era. Cotton demand grew at better than 5 percent per year between 1820 and 1860, a rate of growth never equaled thereafter for such a sustained period. This growth was based on the opening of new markets for British cotton textiles, a phenomenon which could not continue indefinitely at the same pace.[19] Indeed, the first major phase of expansion had largely played out by 1860. When the *U.S.* cotton crop finally recovered to its 1860 level in the crop year 1877–78, the price was exactly the same as it had been before the war: demand had gone nowhere for 18 years, implying at least a 30 percent fall in southern per capita cotton earnings. Since F-E acknowledge that they find an efficiency advantage for slavery only in a handful of commercial crops (1977, p. 292), of which cotton was easily the most important, slavery's dependence on this burgeoning external demand is implied as much by their analysis as by mine. But after the period of postwar recovery, American cotton still dominated world production, and there was no increase in price representing the costs of less efficient production under free labor.[20] The economics of American slavery look very different from this historical perspective.

It seems to me that rather than calculating efficiency indices and then attempting to adjust and refine them to represent special historical and institutional elements of the situation, we would be better off to frankly acknowledge the ambiguities of these measures as applied to slavery, and that by exploring these ambiguities we may come to a better understanding of the economic nature of slavery. For example, all authorities agree that one of the important economic features of slavery was participation of slave women in fieldwork to a much greater extent than free women. But how shall we characterize this phenomenon? Since the returns to female fieldwork seem to have been higher than the alternatives, one might call it an increase in efficiency. Alternatively, one might call it an increase in female "labor force participation" and thus in reality a higher level of input. One might even say, since such activities as cooking and childcare can be centralized, that economies of scale are involved. There is a logic to each of these choices. But it seems to me that the central element is the same as the choice between cotton and corn: slavery involved the involuntary reallocation of family labor from nonmarket economic activity to production of crops for sale. In this alternative conception, it was the *interaction* between the crop mix and female field work which gave slavery its distinctive advantage: per capita food production was roughly similar on farms of all sizes, but this common pattern of self-sufficiency implied an extremely high payoff to the reallocation of "marginal" labor from the household to the fields. When slavery ended, black family labor allocation moved markedly toward a more conventional pattern (see Roger Ransom and Richard Sutch, pp. 22–23). From this point of view, the efficiency of slavery is historically specific to an era in which free households chose to sharply limit their participation in the market economy.[21]

Instead, F-E now interpret the efficiency

[19]See Lars Sandberg's recent history of the Lancashire industry. "This earlier growth had been based principally on the opening-up of new markets. . . It is utterly reasonable to have expected such progress would continue up to World War I, especially with regard to the growth of output and exports" pp. 131, 180).

[20]Cotton prices were high during the transition period 1866–77, but there is no trend in real cotton prices over the whole period 1825–1910:

$$\ln P = 2.25 - \underset{(0.50)}{0.0008t}; \qquad R^2 = .003$$

[21]This theme is developed in my book (1978).

of slavery as a matter of the greater intensity of slave labor time per hour (1977, pp. 291–94). This view is not based on direct evidence: it is virtually the only remaining choice, given that F-E have convinced themselves that slaves worked no more hours than free farmers, and that neither crop mix effects nor distinctive features of the census year need be allowed for. The best of our models are no more than metaphors, and perhaps there is a grain of truth in the claim that an hour of cotton cultivation is a more intensive activity than an hour behind the stove. But if this were the best metaphor, it would be difficult to say why the efficiency advantage should be limited to a handful of southern commercial crops. And if an assembly-line analogy were appropriate, one would be hard pressed to say why the factory was not the ideal setting for slavery. The implication that the "assembly line" and the "speedup" were not well adapted to "urban industries" (see F-E, 1977, p. 292) should be hard for any economic historian to swallow. But if the economic essence of slavery was the shift of family labor from nonmarket to market activity, then it comes as no surprise that the "productivity advantage" disappears whenever output is fully monetized.

REFERENCES

H. Helm Clayton, *World Weather Records*, Smithsonian Miscellaneous Collections, Vol. 79, Washington 1927.

James R. Covert, *Seedtime and Harvest*, U.S. Department of Agriculture, Bureau of Statistics Bull. 85, Washington 1912.

P. A. David and P. Temin, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South: Comment," *Amer. Econ. Rev.* Mar. 1979, *69*, 213–18.

R. H. Day, "Probability Distributions of Field Crop Yields," *J. Farm Econ.*, Aug. 1965, *47*, 713–41.

Ezekiel J. Donnell, *Chronological and Statistical History of Cotton*, New York 1872.

Robert W. Fogel and Stanley Engerman, *Time on the Cross*, Vols. I, II, Boston 1974.

——— **and** ———, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South," *Amer. Econ. Rev.*, June 1977, *67*, 275–96.

R. E. Gallman, "Self-Sufficiency in the Cotton Economy," *Agr. Hist.*, Jan. 1970, *44*, 5–23.

———, "A Note on the Patent Office Crop Estimates, 1841–1848," *J. Econ. Hist.*, June 1963, *23*, 185–95.

P. H. Grissom and W. I. Spurgeon, "Fertility Practices for Cotton and Corn in the Yazoo-Mississippi Delta," Mississippi State College Agr. Experim. Stat. Bull. no. 614, Apr. 1961.

R. Ransom and R. Sutch, "The Impact of the Civil War and of Emancipation on Southern Agriculture," *Explor. Econ Hist.*, Jan. 1975, *12*, 1–28.

Lars Sandberg, *Lancashire in Decline*, Columbus 1974.

Alfred G. Smith, *Economic Readjustment of an Old Cotton State*, Columbia 1958.

W. I. Thorp, *Business Annals*, New York 1926.

James L. Watkins, *King Cotton*, New York 1908.

Gavin Wright, "Prosperity, Progress and American Slavery," in Paul A. David et. al., *Reckoning with Slavery*, New York 1976.

———, "Slavery and the Cotton Boom," *Explor. Econ. Hist.*, Oct. 1975, *12*, 439–51.

———, *The Political Economy of the Cotton South*, New York 1978.

——— **and H. Kunreuther,** "Cotton, Corn and Risk in the Nineteenth Century," *J. Econ. Hist.*, Sept. 1975, *35*, 526–51.

American Cotton Planter, Vol. IV, Montgomery, Apr. 1860.

U.S. Commissioner of Patents, *Annual Reports* for 1849, 1850, Washington.

Monopoly and the Rate of Extraction of Exhaustible Resources: Note

By TRACY R. LEWIS, STEVEN A. MATTHEWS, AND H. STUART BURNESS*

In a recent paper appearing in this *Review*, Joseph Stiglitz demonstrates under a set of familiar conditions that a monopoly-owned nonreplenishable resource will tend to be exhausted at a slower rate than is socially optimal.[1] This supports earlier views on the subject expressed by Harold Hotelling and Robert Solow. Stiglitz shows under the natural "first approximation" assumptions of stationary, iso-elastic demand and zero extraction costs, that monopolistic and socially optimal (competitive) extraction rates are identical. If demand elasticity increases with time or constant unit production costs are positive but possibly decrease with time, he shows that competitive extraction rates exceed monopolistic rates for at least an initial period of time.

In this note we present realistic, alternative extensions to the iso-elastic, zero cost analysis which tend to bias monopolistic extraction rates in the opposite direction, that is, towards excessive resource use. The first modification allows for costs that do not vary with the extraction rate. Occurring in the form of leasing fees, capital costs, and maintenance fees, these quasi-fixed costs[2] are incurred only during periods of production and often constitute a substantial portion of operating expenses.[3]

The second extension involves demand elasticities varying with consumption instead of time. In particular, we consider a stationary demand schedule with elasticity increasing in consumption. A justification for this assumption is that for small quantities demand may be inelastic if certain amounts of the resource are essential in the production of some goods. At lower prices, however, the resource may be used in other industries for which substitute inputs exist as well. Consequently, the elasticity of aggregate demand may increase. For example, the demand for natural gas by homeowners with gas appliances is inelastic since substitute fuels are difficult to use. Since these homeowners are the primary users at high prices, natural gas demand is inelastic at high prices (low consumption). However, at lower prices marginal usage occurs at the extensive margin as various manufacturers switch to natural gas, and aggregate demand may consequently become more elastic.[4] The result in this case is that if costs are quasi fixed, or if demand elasticity increases with consumption, then a monopolist depletes the resource too soon.

In general, competitive ownership of the resource will also result in socially nonoptimal production when fixed operation costs exist. For example, if all costs are quasi fixed, least cost production requires that only one mine operate at a time. Yet with discounting there always will be an in-

*Departments of economics, University of Arizona, University of Illinois-Urbana, and University of New Mexico, respectively. This research was performed at the Environmental Quality Lab, California Institute of Technology, in part under the auspices of Ford Foundation grant #740-0469 and Energy Research and Development Administration grant #EX-76-G-03-1305, Caltech Energy Research Program and is gratefully acknowledged. We wish to thank Jim Quirk and Vernon Smith for thoughtful comments and suggestions on an earlier version of this manuscript. Thanks are also due to George Borts and a referee for suggestions that led to this final form. Of course, remaining deficiencies are our responsibility.

[1]Similar analyses comparing monopolistic and socially optimal extraction rates appear in John Kay and James Mirrlees, Tracy Lewis, James Sweeney, and Milton Weinstein and Richard Zeckhauser. Milton Kamien and Nancy Schwartz compare extraction rates in a general equilibrium setting.

[2]Fixed costs of this variety which can be avoided by stopping production were categorized as "avoidable fixed costs" by Vernon Smith, pp. 257–59.

[3]For example, see James Hendry.

[4]Strong income effects may also tend to cause demand elasticity to increase with consumption.

centive for competitive mining firms to operate simultaneously. While several alternative forms of market intervention might limit the number of operating mines to the social optimum, in general the behavior of an unregulated competitive industry that has quasi-fixed costs is difficult to assess and beyond the scope of this note.[5] Consequently, we contrast monopolistic with socially optimal programs of resource extraction.

Letting $p(q)$ be the inverse demand function for the resource, Q_o be the initial resource supply, and $q_S(t)$ and $q_M(t)$ be the socially optimal and monopolistic rates of extraction, the respective maximization problems for the social maximizer[6] and the monopolist are:[7]

$$(1)\quad \underset{q_S(t),T_S}{\text{maximize}} \int_0^{T_S} \left[\int_0^{q_S(t)} \cdot p(q)dq - F\right]e^{-rt}dt$$

subject to $\int_0^{T_S} q_S(t)dt \leq Q_o; q_S(t), T_S \geq 0$

$$(2)\quad \underset{q_M(t),T_M}{\text{maximize}} \int_0^{T_M} \cdot [p(q_M(t))q_M(t) - F]e^{-rt}dt$$

subject to $\int_0^{T_M} q_M(t)dt \leq Q_o; q_M(t), T_M \geq 0$

where r is the discount rate and T_S and T_M are the terminal extraction dates. Note that these terminal dates are choice variables.

Performing the indicated maximizations, manipulation of the necessary conditions for (1) and (2) yields, respectively;

$$(3)\quad \dot{q}_S(q) = \frac{rp(q)}{p'(q)} = -re(q)q$$

$$(4)\quad \dot{q}_M(q) = \frac{rR'(q)}{R''(q)} = -re(q)q\left[1 - \frac{e'(q)q}{e(q)-1}\right]^{-1}$$

where $R(q) = p(q)q$ is the revenue function, which we presume is concave, and $e(q)$ is the demand elasticity. We assume $e > 1$ to ensure positive monopolistic output so that (3), (4), and $e'(q) \geq 0$ imply[8]

$$(5)\quad 0 > \dot{q}_S(q) \geq \dot{q}_M(q)$$

The necessary terminal time conditions can be expressed as[9]

$$(6)\quad q_M(T_M)f'(q_M(T_M)) = f(q_S(T_S)) = F$$

where the function $f(q)$ is defined by

$$(7)\quad f(q) \equiv \int_0^q p(x)dx - qp(q)$$

Note that $f''(q) \equiv p'(q)e(q)^{-1} - p(q)e'(q)e(q)^{-2} < 0$ since $e'(q) \geq 0$. The concavity of f together with $f(0) = 0$ and (6) imply[10]

$$(8)\quad q_M(T_M) \geq q_S(T_S)$$

Changing variables of integration from t to q in the resource constraint equations yields

$$(9)\quad -Q_o = \int_{q_M(T_M)}^{q_M(0)} [q/\dot{q}_M(q)]dq = \int_{q_S(T_S)}^{q_S(0)} [q/\dot{q}_S(q)]dq$$

Consistency between (5), (8), and (9) requires that $q_M(0) \geq q_S(0)$; i.e., the monopoly initially extracts at a rate no slower than is socially optimal. Since inequality (5) is strict if $e'(q) > 0$, and inequality (8) is strict if $F > 0$, the initial monopoly extraction rate will be excessive in either case. From equation (5) the time path $q_M(t)$ crosses $q_S(t)$ at most once, and only from above. Thus the monopolist either extracts

[5]This topic is currently being pursued by us in a subsequent manuscript.

[6]Subject to the usual caveats, the social maximizer is assumed to maximize consumer's surplus, the area beneath the demand curve.

[7]Since cost minimization with zero variable and positive quasi-fixed costs requires that one mine operate at a time, F represents the fixed operating costs for one mine.

[8]Lest the point of this section be made vacuously we hasten to assert that demand functions satisfying these requirements exist. In particular if the social welfare function is $U(q) = ln\ q + 2q^{1/2}$ then $dU/dq = q^{-1} + q^{-1/2} = p(q)$. From this one easily obtains $e > 1$ and $e'(q) > 0$. Moreover $R''(q) < 0$ everywhere.

[9]Terminal time conditions can be obtained by maximizing the Lagrange expression for this problem with respect to T_S and T_M.

[10]If $F = 0$, we have $T_M = T_S = \infty$ and $q_M(\infty) = q_S(\infty) = 0$. This follows because $e'(q) \geq 0$ implies $\lim_{q \to 0} p(q) = \infty$

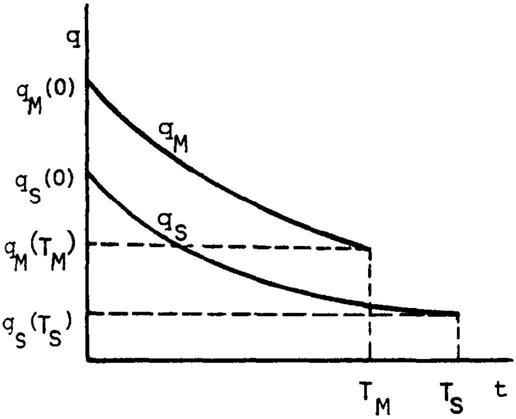


FIGURE 1. NONINTERESECTING EXTRACTION PATHS (ISO-ELASTIC DEMAND AND POSITIVE FIXED COSTS)

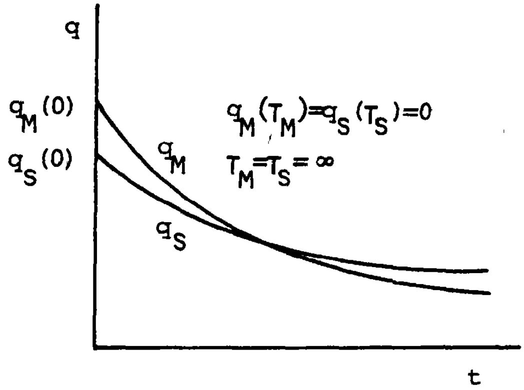


FIGURE 2. INTERSECTING EXTRACTION PATHS (FOR EXAMPLE, ZERO FIXED COSTS AND DEMAND ELASTICITY INCREASING WITH QUANTITY)

too fast for the entire extraction period before exhaustion (see Figure 1), or too fast initially and too slowly thereafter (see Figure 2). Hence, the resource remaining is always less than socially optimal and depletion occurs too soon.[11] This is clear for the case in Figure 1, and the case in Figure 2 follows directly from $q_M(0) > q_S(0)$, $q_M(\infty) = q_S(\infty) = 0$ and the fact that the paths $q_M(t)$ and $q_S(t)$ intersect only once.

Thus we have established the following:

PROPOSITION 1: *Suppose that demand for a nonrenewable resource is stationary and everywhere elastic, and that all variable costs are zero. If either* (a) *quasi-fixed costs are positive and demand elasticity is nondecreasing in consumption, or* (b) *quasi-fixed costs are nonnegative and demand elasticity is strictly increasing in consumption, then a monopolist extracts the resource faster than is socially optimal in the following ways:* (i) $T_M \leq T_S$ *(with* $T_M < T_S$ *if* $F > 0$), (ii) $q_M(t) > q_S(t)$ *initially (and for all* $t \leq T_M$ *if* $F > 0$ *and* $e' = 0$), *and* (iii) $Q_M(t) < Q_S(t)$ *for* $t < T_S$.

At first glance the excessive extraction rate of the monopolist when $e'(q) > 0$ appears mysterious, if only because derivatives of elasticities are second-order demand characteristics and do not affect usual (static) marginal analyses. But consider a simple two-period world, and suppose the monopolist is considering the socially optimal schedule (q_S^1, q_S^2) that is determined by setting the first period price equal to the discounted second period price. Marginal

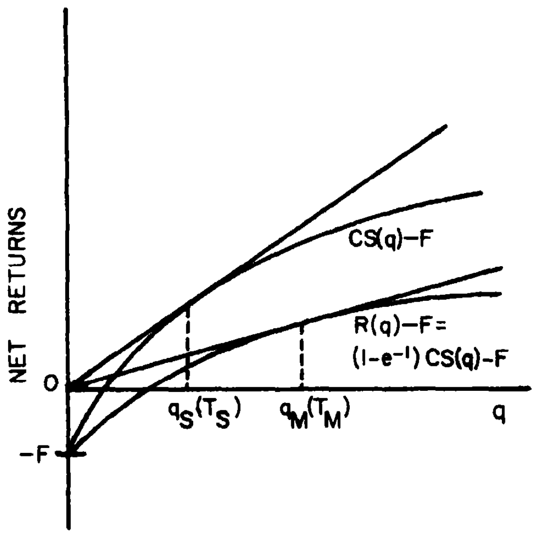


FIGURE 3. MONOPOLISTIC AND SOCIALLY OPTIMAL TERMINAL NET RETURNS AND OUTPUT (POSITIVE QUASI-FIXED COSTS AND ISO-ELASTIC DEMAND)

[11]For $F = 0$, we obtain the general result $Q_M(t) \lesseqgtr Q_S(t)$ as $e'(q) \gtreqless 0$. The analysis for $e'(q) < 0$ is in Lewis.

revenue is a fraction, $1 - e^{-1}$, of price, and that fraction increases with demand elasticity. Since $q_S^1 > q_S^2$ implies $e(q_S^1) > e(q_S^2)$, equality of the two discounted prices implies that discounted marginal revenue is greater in the first period than in the second. The monopolist therefore adjusts (q_S^1, q_S^2) by extracting more in period one and less in period two, until the two discounted marginal revenues are equal. When the number of periods is variable, this same reasoning indicates that the monopolist depletes the resource too soon.

The other polar case of positive quasi-fixed costs and iso-elastic demand can also be understood more heuristically. The instantaneous net returns of the monopolist and social maximizer are $(1 - e^{-1})\ CS(q) - F$ and $CS(q) - F$, respectively, where $CS(q)$ is consumer surplus. Both the monopolist and social maximizer bear the same costs F to operate in each period, but the monopolist captures but a fraction of the returns accruing to the social maximizer. There is, thus, a greater incentive for the monopolist to accelerate extraction to reduce total operating costs FT_m (see Figure 3).

We have shown that in two special cases a monopolist depletes a natural resource faster than is optimal. Since Stiglitz proves the opposite result for other special cases, the net effect of all these presumably realistic considerations is analytically indeterminate and must be ascertained empirically.

REFERENCES

J. B. Hendry, "The Bituminous Coal Industry," in Walter Adams, ed., *Structure of American Industry*, New York 1961.

H. Hotelling, "The Economics of Exhaustible Resources," *J. Polit. Econ.*, Apr. 1931, *39*, 137–75.

M. I. Kamien and N. L. Schwartz, "The Optimal Resource-Capital Ratio and Market Structure," *J. Econ. Theory*, forthcoming.

J. A. Kay and J. A. Mirrlees, "The Desirability of Natural Resource Depletion," in David W. Pearce and James Rose, eds., *The Economics of Natural Resource Depletion*, London 1975, 140–76; 218–20.

T. R. Lewis, "Monopoly Exploitation of an Exhaustible Resource," *J. Environ. Econ. Manage.*, No. 3, 1976, *3*, 198–204.

Vernon L. Smith, *Investment and Production*, Cambridge 1961.

R. M. Solow, "The Economics of Resources or the Resources of Economics," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 1–14.

J. E. Stiglitz, "Monopoly and the Rate of Extraction of Exhaustible Resources," *Amer. Econ. Rev.*, Sept. 1976, *66*, 655–61.

J. Sweeney, "Economics of Depletable Resources: Market Forces and Intertemporal Bias," *Rev. Econ. Stud.*, Feb. 1977, *44*, 125–42.

M. Weinstein and R. J. Zeckhauser, "The Optimal Consumption of Depletable Natural Resources," *Quart. J. Econ.*, Aug. 1975, *89*, 371–92.

Monopoly and the Rate of Extraction of Exhaustible Resources: Note

By GORDON TULLOCK*

Surely there are very few people in the world who doubt "that the oil producing countries ... have been acting collusively and have forced the price of oil to a level far higher than it would have been in competitive equilibrium" (see Joseph Stiglitz, p. 655). In the face of what has happened in the real world, few would agree "that there is a very limited scope for the [oil] monopolist to exercise his monopoly power" (p. 655).

In fact, not only have the oil producing countries made a great deal of money from their cartel, but this is what theory indicates should happen. The error in the Stiglitz article can very readily be seen by examining the "intuitive" model used on the first page. This involves a two-period model with zero extraction costs and a constant elasticity of demand. There is also a quite unrealistic assumption: "That part of the stock which we do not consume in the first period will be consumed in the second" (p. 655). This assumption means that there is no possibility of any monopoly profit because this requires, ex definition, that the monopolist sell less than the competitive volume. If we permit the monopolist to waste part of the oil in this particular model and assume that the demand is inelastic, then it is obvious that the monopoly with "constant [in]elasticity demand schedules" would sell one drop of oil in each of the two periods and waste the remainder. This would lead to higher prices and a higher revenue for the monopolist than would be obtained in the competitive market. If we assume that the "constant elasticity demand" is an *elastic* demand (and this does not fit the oil case), then the combination of the zero cost of extraction and the constant elasticity demand schedule *does* lead the monopoly to sell its oil at the same price and rate as under competition.

The constant elasticity assumption is usually used for mathematical convenience, but, so far as I know, no one thinks that it fits any real world situation. In the real world, the elasticity varies a good deal, depending on where you are on the demand curve, and these variances affect the optimal monopoly profit.

The problem of elasticity is discussed in a footnote of the Stiglitz article: "Some have suggested that the demand for oil in the very short run has less than unitary elasticity,..." (p. 656, fn. 2). As a matter of fact, the empirical calculations for elasticity of oil before the cartel was formed showed an elasticity of approximately 0.15, and the actual response of sales and prices would seem to indicate that this was a good estimate. The transfer of resources to the members of the cartel in the first year of its foundation was of the order of $50 billion, which, in many cases, meant that the individual countries' revenue rose by several hundred percentage points.

Continuing to use the very simple intuitive model of the first page of the Stiglitz article, let us inquire what would happen if the demand were inelastic in the first period and became elastic in the second. Under these circumstances, the oil monopoly would sell one drop of oil in the first period and all the rest in the second period, if it is compelled to do so by the Stiglitz rule that it must exhaust supplies. This would lead to higher net revenues than would be obtained by the cartel. Note also that it sharply changes the time distribution of the consumption of oil in the sense that it is deferred. It is true, however, that the change is, as Stiglitz says, dynamically inefficient.

This is a general principle. A monopoly

*Center for Study of Public Choice, Virginia Polytechnic Institute and State University.

of an exhaustible resource will normally produce less in the current period, and therefore have more left over for future periods, than would a competitive market. However, this is not because the monopolist has a different evaluation of current and future revenues or because he is consciously saving the resource for the future. It results simply from the fact that the monopolistic restriction of present-day production is most conveniently achieved by simply leaving the oil in the ground rather than by taking it out and wasting it in some way. He conserves the oil rather than taking it out of the ground and destroying it, simply because that is a cheaper way of lowering his production and thereby raising his prices.

This does mean that he will have more oil next year than he would if he had followed competitive pricing levels, but, since he is planning to sell less oil next year than the competitive market would anyway, the value of this surplus to him is presumably very small. The monopolist conserves resources as a by-product of raising prices in the current period, not because he wants to hold the resource for sale later.

Stiglitz says that

> The basic argument is a simple one: the monopolist, like the competitor, eventually will exhaust all of the natural resource. It is not like a conventional commodity, where the total amount that will eventually be sold is smaller for a monopolist than for a competitor.... Since equilibrium entails exhaustion of the stock of resources as time approaches infinity, the competitive market equilibrium and the monopoly are described by exactly the same set of equations: the two equilibria are identical. [p. 656]

Suppose that the demand for oil will always be inelastic. Further, assume that we keep the zero extraction costs. In every year from now to infinity, the monopolist could charge a higher price and sell fewer units of oil than the competitive market, and would make a profit doing so. Further, in every year from now +1 to infinity, the reserves left in the ground would be larger under the monopoly than under competitive conditions. The fact that mathematically the infinity time period leads to odd results is unfortunate, but it does not change the above situation. The monopolist would be keeping the oil in the ground each year, not because he expects to sell it for more next year but because it affects the price this year.[1]

With more reasonable assumptions about elasticity and cost, it seems quite probable that the long-run demand elasticity of oil is above 1, although, granted its superiority as a fuel, this may not be so in the present range of prices. In the long run, however, the cost of extraction ceases to be zero and becomes very substantial. Under the circumstances, a monopoly, balancing marginal revenue with marginal cost in the long run, once again will produce a lower volume and get a higher price than a competitive market balancing price and the marginal cost. The biggest problems essentially come from Stiglitz's implicit assumption that the reason the monopolist would withhold oil from production is a feeling that it is worth more in the future than it is now, rather than that he withholds oil from production as a by-product of attempting to get a higher price now. To take an extreme example, suppose the monopolist uses a higher price and lower production *this year*. This will mean that he will have more on hand next year than he would if he had followed competitive rules. It does *not* mean, however, that next year he is compelled to sell more than would be sold in a competitive market. From this period on, he can simply sell exactly the same amount and at the same price as the competitive market, and hence he would retain the profit on the first period and suffer no loss on later periods. This is more profitable than selling

[1]Note that this is not necessarily a profit-maximizing course of action. When the supplies of oil under the competitive market were very low, and therefore the price was very high, the monopolist—finding himself with much larger supplies in the ground as an accidental by-product of his earlier profit-maximizing activity—might in fact choose a lower price, a higher sale volume, and much higher total revenue than the competitive market would achieve at this late stage.

this year at competitive prices, even if it is not strictly speaking a profit-maximizing course of action.

The oil cartel is not alone. Indeed, it is not even the first oil cartel. Nickel, tin, platinum, and diamonds are all mineral resources with presumably finite total supplies, and all at one time or another have been cartelized. Stiglitz implicitly argues that the businessmen who thought they saw immense profits, and the stockholders who later thought they had received immense profits from these cartels, were wrong.

REFERENCES

J. E. Stiglitz, "Monopoly and the Rate of Extraction of Exhaustible Resources," *Amer. Econ. Rev.*, Sept. 1976, *66*, 655–61.

Monopoly and Crude Oil Extraction

By JOHN J. SOLADAY*

Recently, increasing attention has been directed toward analyzing the impact of monopoly control on the utilization of depletable resources. Several theoretical works have treated the problem with varying degrees of generality. This literature indicates that under most circumstances approximating reality, the monopolist will overconserve resources relative to the competitive optimum. However, there does not seem to be general agreement on the degree of severity of the intertemporal bias caused by the existence of monopoly. The purpose of this paper is to develop some quantitative evidence of the differences between competitive and monopolistic behavior. By comparing estimated production profiles of oil producing states grouped according to their degree of market power, estimates of the extent to which relatively competitive and monopolistic output profiles differ are acquired. Results indicate that the monopoly type group exhibits a production profile which is initially lower, peaks later, and eventually exceeds the profile estimated in the competitive group.

The owner of a resource, who is motivated by a desire to maximize its present value in a competitive market, will schedule production so that the difference between price and marginal cost will increase over time at the rate of interest.[1] If this condition were not met, it would be possible to increase wealth by shifting output among time periods. In a monopoly situation, the equilibrium condition for present value maximization requires that the difference between marginal revenue and marginal cost must rise over time at the rate of interest (see Joseph Stiglitz and James Sweeney). However, the extent to which competitive and monopoly profiles differ will depend upon specific conditions relating to the behavior of demand and cost functions over time.

The paper by Milton Weinstein and Richard Zeckhauser establishes that a competitive market will result in a consumption stream which maximizes consumer's plus producer's surplus and can be considered optimal. In their zero extraction cost model, the market price will also rise by the interest rate. A monopolist's production profile will coincide with a competitor's under conditions of constant and stable elasticity of demand. If price elasticities increase with time, present value maximization dictates that marginal revenue increase with the interest rate. Price will increase by less than the interest rate, and the monopolist will be overconserving from a social standpoint.

Stiglitz' analysis also concludes that competitive and monopolistic production profiles will coincide under conditions of zero extraction costs and constant elasticity of demand. Under the more realistic conditions of increasing elasticity of demand, the monopolist will adopt a more conservationist policy by producing in a manner similar to Figure 1. Stiglitz also concludes that a similar bias arises when extraction costs are accounted for in the constant elasticity of demand case.

The Sweeney study examines the intertemporal biases resulting from several market imperfections. The impact of depletion allowances, monopoly, externalities, price regulation, and international vulnerability, are determined through the creation of a market imperfections function.

In the case of monopoly, Sweeney estab-

*Economist, Exxon Corporation. This article was written while I was an assistant professor of economics, Pennsylvania State University. I would like to thank Richard Gordon, Yash Mehra, James Sweeney, and Willard Witte for helpful comments.

[1] In the case in which costs increase with cumulative production, an even slower growth of the difference between price and marginal cost is required for present value maximization. See Ronald Cummings.

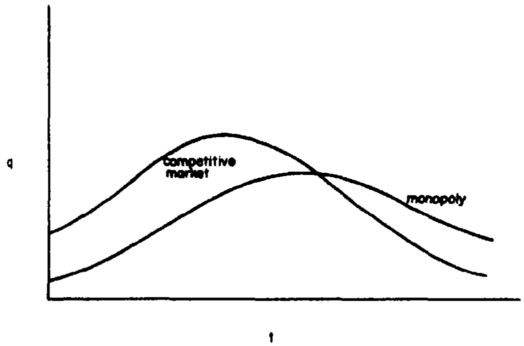


FIGURE 1. THE PRODUCTION PROFILES

lishes the above mentioned results. In addition he treats more realistic demand and cost functions simultaneously. His work indicates that as long as the demand elasticity is not decreasing over time and there are non-time-increasing extraction cost functions,[2] a monopolist will produce at a lower rate than will a perfect competitor. Since both of these conditions are likely to occur, the monopolist can be expected to overconserve resources.

The oligopolistic situation in the *U.S.* oil industry provided only a rough testing ground for the hypotheses contained in the above mentioned papers. By grouping states according to the type of control on oil production, this paper attempts to ascertain the quantitative difference between production under conditions resembling monopoly and competition. The oil producing states in this country could have been divided into three institutionally distinct groups depending upon the type of oil production control adopted. Five "market demand" states[3] approximated a monopoly situation. These governmental regulatory structures had arisen ostensibly to prevent physical waste originating from uncontrolled oil production, but had effectively maintained above market-clearing prices and monopoly rents. Essentially, state agencies in the market demand states limited oil production to industry forecasts of quantities of oil demanded for each state (at presumably, but nonspecified, prices equal to or higher than previous levels). These forecasted quantities (net of predicted unregulated production) were subsequently allocated among wells by permitting regulated wells to produce state determined percentages of an assigned maximum allowable production level. The percentages were uniform across wells and varied over time to insure that production just equaled state forecasts.[4] A second group of six states had no direct controls on oil production and are considered to represent a competitive group.[5] Finally, an intermediate group is comprised of seven states which controlled oil production with the aim of preventing it from exceeding technologic limits (maximum efficient rates, *MER*) essentially designed to maximize the undiscounted stream of current and future oil recovery from reserves.[6]

In order to determine what the differences in the production time paths were among the state groups, individual state data on reserves, production, and yearly gross additions to reserves (new oil) were acquired from the American Petroleum Institute. The data sample consisted of the period 1948–74 for the eighteen states which accounted for over 97 percent of *U.S.* production during the period. In order to adjust for size differences among states, all data are divided by state reserves as of 1960.

The data grouping cited here is the state cross section pooled over time. State time-series and overall regressions were also estimated. However, their generally similar coefficient estimates and their poorer performance in terms of higher standard errors of estimate lead us to exclude reporting

[2] In addition, the marginal revenue function is assumed not to grow at a rate greater than the interest rate.

[3] The market demand states consist of Texas, Louisiana, New Mexico, Oklahoma, and Kansas.

[4] The market demand proration system is more fully discussed in Paul Homan and Wallace Lovejoy; Interstate Oil Compact Commission; also see Stephen McDonald.

[5] The unregulated group is comprised of Illinois, Indiana, Ohio, Pennsylvania, Kentucky, and West Virginia.

[6] The *MER* group includes Arkansas, California, Colorado, Mississippi, Montana, Nebraska, and Wyoming.

them.[7] The basic data matrix consists of observations of the eighteen states for each of twenty-seven years. The states were put into groups of six, seven, and five corresponding to the unregulated, *MER*, and market demand group described above. Ordinary least squares (*OLS*) regressions were performed by aggregating product moments of the data in order to acquire weighted average regression coefficients. The product moments, as described by Robert Eisner, consist of differences between individual state observations for each year and the mean of all state observations within each group calculated on a yearly basis. These deviations are pooled over all years in the sample. These results then represent an "average" of individual cross-section regressions.

The production profile was ascertained by use of the rational lag estimator where oil production (q_t) is a distributed lag function of current and past contributions to the stock of developed oil reserves (n_t). The sequence of the w's in equation (1) represents the profile of structural coefficients estimated by the rational lag technique.

$$(1) \quad q_t = w_o n_t + w_1 n_{t-1} + w_2 n_{t-2} + \dots + w_n n_{t-n}$$

Technological information indicates that production from crude oil reserves usually builds up for the first few periods after new oil is reported. The Koyck scheme of declining geometric weights was, therefore, not imposed until the third period of production. The results of the estimating equation are presented in Table 1. Estimates of

TABLE 1—OIL OUTPUT AS A FUNCTION OF CURRENT AND LAGGED NEW OIL AND LAGGED OUTPUT[a]
(By State Group, Pooled Cross-Section Regressions)

Variable or Statistic	Regression Coefficient and Standard Errors: Unregulated	*MER*	Market Demand
Constant	.0000	.0000	.0000
	(.0011)	(.0008)	(.0003)
n_t	.0693	.0462	.0432
	(.0140)	(.0100)	(.0086)
n_{t-1}	.0157	.0529	.0240
	(.0146)	(.0104)	(.0090)
n_{t-2}	−.0111	.0163	.0311
	(.0145)	(.0107)	(.0094)
q_{t-1}	8749	.8424	.8924
	(.0344)	(.0297)	(.0156)
Σn	.0739	.1154	.0982
	(.0230)	(.0148)	(.0120)
N[a]	144	168	120
R^2	.9089	.8928	.9840
F	346.7	339.3	1767.4
SEE	.0134010	.0101758	.00343980
$\alpha = \frac{\sum_{i=o}^{2} w^i}{1-b_1}$	.5907	.7322	.9136
	(.2567)	(.1984)	(.2222)

[a] N = number of observations. The estimating equation is

$$q_t = \sum_{i=o}^{2} w_i n_{t-i} + b_1 q_{t-1}$$

the sum of the production profile coefficients (α) or recovery rates from reported new oil were .59, .73, and .91 in the unregulated, *MER*, and market demand states. The production profiles themselves are reported in Table 2. The production profile is lengthened in the market demand states with the ratio of cumulative production for the first sixteen years to eventual production ($\Sigma_{i=o}^{15} w_i/\alpha$) equalling only .82 whereas the ratio was .88 in the unregulated and .93 in the *MER* states. The median lag on the production was six years in the unregulated states, five years in the *MER*, and eight years in the market demand states.

The description of crude oil reserves data indicate the recovery rates should approximate unity. I suspect that an errors in variables problem, arising because of differences between reported and economically producible new oil, could have biased downward my estimates of the production

[7]The lower standard errors of the pooled cross-section regressions are interpreted as indicating relatively greater similarity of the production time paths within state groups for each time period. The estimated relationship was more stable within each of the time periods (cross sections) than across time periods (time-series). I can only conjecture about this result. It may be that disturbances that occurred over time produced greater bias or unexplained disturbance than variables that changed among states. My original hypothesis was that the institutional characteristics were the important variables in determining differences among the states. The better cross-section performance provides at least some support for this conjecture.

TABLE 2—THE PRODUCTION PROFILES DERIVED FROM TABLE 1[a]

Structural Coefficient	Unregulated	*MER*	Market Demand
w_0	.0693	.0462	.0432
w_1	.0764	.0918	.0626
w_2	.0557	.0937	.0869
w_3	.0487	.0789	.0776
w_4	.0426	.0665	.0692
w_5	.0373	.0560	.0618
w_6	.0326	.0472	.0551
w_7	.0286	0.397	.0492
w_8	.0250	.0335	.0439
w_9	0219	.0282	.0392
w_{10}	.0191	.0238	.0350
w_{11}	.0167	.0200	.0312
w_{12}	.0146	.0169	.0278
w_{13}	.0128	.0142	.0248
w_{14}	.0112	.0120	.0222
w_{15}	.0098	.0101	.0198
$\sum_{i=0}^{15} w_i$	.5224	.6786	.7495
α	.59097	.73249	.91357
Median Lag	5.1	5.8	7.0
Σ/α	.8840	.9264	.82036

[a]The structural equation is

$$q_t = w_0 n_t + w_1 n_{t-1} + w_2 n_{t-2} + \ldots + W_{15} n_{t-15}$$

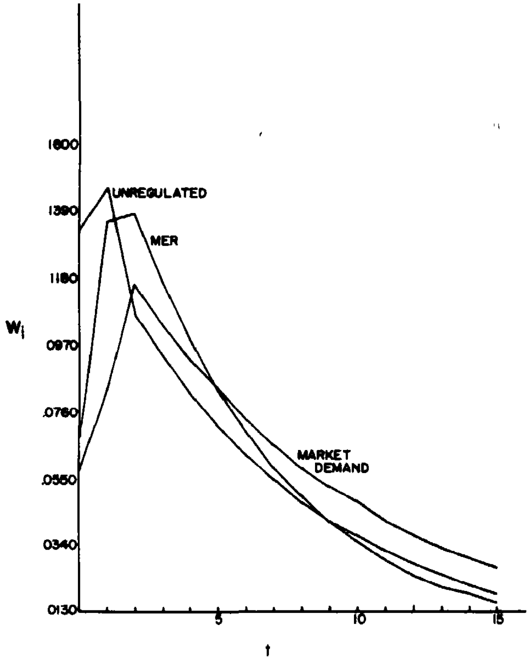


FIGURE 2. MODIFIED KOYCK ESTIMATES OF THE PRODUCTION PROFILES

profile. In an attempt to adjust for this, all production profile weights, acquired from Table 2, were blown up to sum to unity, that is, individual weights were divided by the recovery rate estimates. The resulting profiles are shown in Figure 2. It should be noted that the imposition of a unitary recovery rate is consistent with the assumption in both the Stiglitz and Sweeney papers that the monopolist and competitive firms by selecting different production profiles will not affect total recovery. In certain cases, however, recovery rates could be dependent upon the rapidity of extraction. In an attempt to determine the accuracy of inflating the production time path weights to sum to unity, a predicted output series was constructed according to equation (2):

$$(2)\quad q_t^p = w_0' n_t + w_1' n_{t-1} + w_2' n_{t-2} + \ldots + w_{15}' n_{t-15}$$

where $\Sigma_{i=0}^{15} w_i' = 1$.

Since the original data sample covered 1948–74, and the length of the lag in the structural equation lag was fifteen periods, predicted and reported output was compared from 1964–74. The ratio of the sum of predicted output to the sum of reported output was .99, .92, and .84 in the unregulated, *MER*, and market demand states. Since the original production weights (see Figure 2) were already inflated to sum to unity, we expected the predicted and reported sums to be approximately equal. For every barrel of oil reported in the unregulated group 1.01 (1/.99) barrels are eventually produced, however, 1.09 and 1.19 barrels are eventually produced in the *MER* and market demand groups. This indicates that reserves are underreported by 1, 9, and 19 percent respectively in these groups.

In order to determine whether our results were sensitive to the specific estimating equation used above, several additional regressions were run using the same original data but allowing all combinations of lags up to five periods on both output and new oil. The estimating equations reported in Table 3 were selected based on the mini-

TABLE 3—OIL OUTPUT AS A FUNCTION OF CURRENT AND LAGGED NEW OIL AND LAGGED OUTPUT[a]

Variable or Statistic	Regression Coefficients and Standard Errors: Unregulated	*MER*	Market Demand
Constant	.0000	.0000	.0000
	(.0012)	(.0007)	(.0003)
n_t	.0781	.0499	.0297
	(.0159)	(.0097)	(.0095)
n_{t-1}	.0235	.0407	.0044
	(.0170)	(.0106)	(.0100)
n_{t-2}	−.0024	.0117	.0218
	(.0162)	(.0109)	(.0096)
n_{t-3}	.0343		
	(.0165)		
n_{t-4}			
n_{t-5}			
q_{t-1}	.6698	1.2410	1.1830
	(.0912)	(.0821)	(.0994)
q_{t-2}	−.0021	−.6189	−.1894
	(.1085)	(.1190)	(.1566)
q_{t-3}	.0349	.2169	−.0408
	(.1081)	(.0734)	(.1578)
q_{t-4}	.0122		.1913
	(.0833)		(.1752)
q_{t-5}			−.2322
			(.1043)
Σn	.1383	.0982	.0559
	(.0291)	(.0161)	(.0135)
Σq	.8243	.8170	.9121
	(.0397)	(.0308)	(.0210)
N	126	147	105
R^2	.9109	.9077	.9869
F	149.425	299.409	903.618
SEE	.0134967	.00947269	.00309612
α	.7872	.5364	.6352
	(.3154)	(.1507)	(.2601)

[a]The estimating equation is

$$q_t = \sum_{i=0}^{m} a_i n_{t-i} + \sum_{i=1}^{k} b_i q_{t-i}$$

TABLE 4—THE PRODUCTION PROFILES DERIVED FROM TABLE 3[a]

Structural Coefficient	Unregulated	*MER*	Market Demand
w_0	.0781	.0499	.0297
w_1	.0758	.1027	.0395
w_2	.0482	.1082	.0629
w_3	.0692	.0816	.0657
w_4	.0498	.0565	.0699
w_5	.0358	.0431	.0683
w_6	.0269	.0362	.0678
w_7	.0205	.0305	.0623
w_8	.0155	.0248	.0562
w_9	.0117	.0198	.0488
w_{10}	.0089	.0158	.0416
w_{11}	.0067	.0127	.0339
w_{12}	.0051	.0103	.0265
w_{13}	.0038	.0084	.0195
w_{14}	.0029	.0067	.0133
w_{15}	.0022	.0054	.0078
$\sum_{i=0}^{15} w_i$	.4614	.6127	.7138
$\alpha = \sum_{i=0}^{\infty} w_i$	.4682	.6355	.6337
Median Lag	3.5	3.7	5.7
Σ/α	.9855	.9642	1.1264

[a]The structural equation is

$$q_t = w_0 n_t + w_1 n_{t-1} + w_2 n_{t-2} + \ldots + w_{15} n_{t-15}$$

mum standard error criteria (see Henri Theil, p. 543). The production profiles derived from the estimates reported in Table 3 are presented in Table 4.

The results of the unconstrained estimates are similar to those obtained in the Koyck equations. The median lag of production is still two years shorter in the unregulated states than in the market demand states. The reason that the unconstrained median lags were shorter than the Koyck estimates is apparently due to the build-up period which was possible in the unconstrained estimates. As in the case of the initial Koyck equations, the structural coefficients of Table 4 were inflated to sum to unity. The paths are exhibited in Figure 3 (production profile coefficients were acquired from Table 4 and were inflated to sum to unity).

The production profiles exhibited in Figure 3 correspond to the initial estimates (see Figure 2) and to the postulated path (see Figure 1). The market demand group initially produces at levels below both the unregulated and *MER* group. The market demand states eventually produce at higher levels since their reserves had been depleted more slowly.

A predicted output series was constructed using the unconstrained coefficients which were inflated to sum to unity (see Figure 3). The ratios of the sum of predicted output to the sum of reported output were .92, .94, and .84 in the unregulated, *MER*, and market demand states. This result indicates that reserves were underreported by 9, 6, and 19 percent, whereas the predicted series calculated from the Koyck equations in-

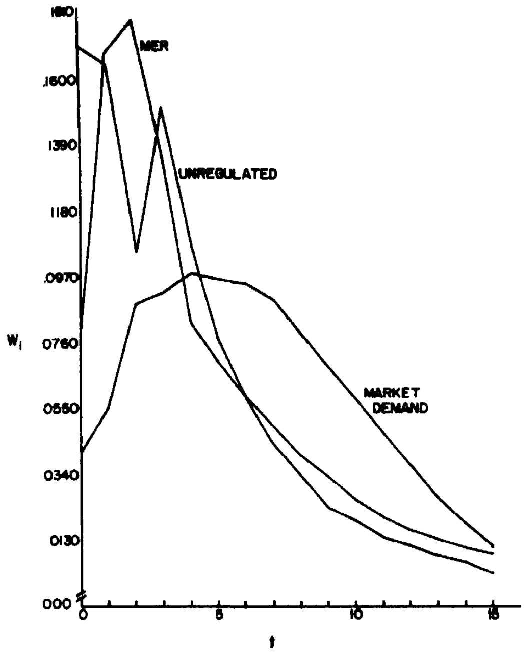


FIGURE 3. UNCONSTRAINED ESTIMATES OF THE PRODUCTION PROFILES

dicated underreporting of 1, 9, and 19 percent in the unregulated, *MER*, and market demand groups. While these estimates are somewhat sensitive to the specific set of production profile coefficients in the unregulated and *MER* groups, the estimate of degree to which reserves are underreported in the market demand group remained the same.

The empirical results contained in this paper indicate that the states with market power selected a production profile that was tilted more toward the future than a competitors' profile. Throughout most of the sample period this has served to increase domestic oil prices above world prices. The monopolistic constraint on oil production was not accompanied by an increase in ultimate recovery. Comparisons of predicted and reported output series indicate that recovery rates are not below unity for any of the state groups. In fact, it appears that the market demand states had underreported reserves by about 19 percent over the period.

One of the problems in this study is the estimation of a fixed production time path for oil. Given the definition of new oil and the production technology of essentially producing these minerals from a working inventory, I suspect that very little is fixed in the true production time paths. Production does respond to changing demand and cost conditions. In this study, responses to changing demand or cost conditions were assumed to be realized by acquiring more or less new oil which was then produced in a predetermined and fixed fashion. The production time paths analyzed in this paper are rough estimates at best, and I can only hope that they approximated the true shifting paths with some accuracy.

REFERENCES

R. Cummings, "Some Extensions of the Economic Theory of Exhaustible Resources," *Western Econ. J.*, Sept. 1969, *7*, 201–10.

R. Eisner, "A Permanent Income Theory for Investment: Some Empirical Exploration," *Amer. Econ. Rev.*, June 1967, *57*, 363–90.

Paul Homan and Wallace Lovejoy, *Economic Aspects of Oil Conservation Regulation*, Baltimore 1967.

Stephen McDonald, *Petroleum Conservation in the United States: An Economic Analysis*, Baltimore 1971.

J. Stiglitz, "Monopoly and the Rate of Extraction of Exhaustible Resources," *Amer. Econ. Rev.*, Sept. 1976, *66*, 655–61.

J. Sweeney, "Economics of Depletable Resources: Market Forces and Intertemporal Bias," *Rev. Econ. Statist.*, Feb. 1977, *44*, 125–41.

Henri Theil, *Principles of Econometrics*, New York 1971.

M. Weinstein and R. Zeckhauser, "Optimal Consumption of Depletable Resources," *Quart. J. Econ.*, Aug. 1975, *89*, 371–92.

American Petroleum Institute, American Gas Association, *Reserves of Crude Oil, Natural Gas Liquids and Natural Gas in the U.S. and Canada and U.S. Productive Capacity as of December 31, 1975*, Washington, May 1976.

Interstate Oil Compact Commission, *A Study of Conservation of Oil and Gas in the United States*, Oklahoma City 1964.

Constant-Utility Index Numbers of Real Wages: Revised Estimates

By JOHN H. PENCAVEL*

In the March 1977 issue of this *Review* I presented a critique of the published Bureau of Labor Statistics (*BLS*) series on real wages by contrasting them with some constant-utility index numbers of real wages. The latter are derived first by solving the representative consumer-worker's indirect utility function for that wage rate which restores some base period's utility after all commodity prices have changed, and second by expressing this constant-utility wage rate as a fraction of actual wage rates. In making these calculations I made use of recently published estimates of the Stone-Geary utility function by Michael Abbott and Orley Ashenfelter. Unfortunately, I subsequently discovered that Abbott and Ashenfelter had miscoded some of their data so that their estimates were incorrect. This also rendered my constant-utility index numbers incorrect. Abbott and Ashenfelter have now reestimated their system of equations and with their revised estimates I now present my recalculated constant-utility real wages.[1] For details of the underlying argument, the reader is referred to the original article.

The recalculated estimates of the parameters of the Stone-Geary function are presented in the first two columns of Table 2. The estimate of γ_h (namely, 2,331 hours per year) implies that the constraint $(\gamma_h - h) > 0$ is not satisfied for the years 1929–33, 1937, and 1939–45. For this reason I do not present estimates of these constant-utility wage rates for these years and, in particular, I have selected 1946 (rather than 1939 as in my earlier paper) as the base year for my constant-utility wage rate calculations.

TABLE 1—PUBLISHED INDEX NUMBERS OF REAL WAGES

Year	(i)[a]	(ii)[b]	(iii)[c]	(iv)[d]
1946	1.000	1.000	1.000	1.000
1950	1.088	1.167	1.051	1.075
1955	1.251	1.380	1.151	1.182
1960	1.383	1.553	1.198	1.246
1965	1.522	1.800	1.330	1.408
1967	1.573	1.866	1.323	1.405

Source: U.S. Bureau of Labor Statistics and Department of Commerce data.

[a] The ratio of average hourly earnings of production workers (as published by the *BLS*) to the consumer price index.

[b] A Laspeyres real wage index based on Department of Commerce data (see original article for its construction).

[c] The *BLS* index of real spendable (i.e., after tax) weekly earnings for a production worker with three dependents.

[d] The *BLS* index of real spendable (i.e., after tax) weekly earnings for a production worker with no dependents.

The first column of Table 3 presents the recalculated series for w_t^*: for any year t, w_t^* gives the wage rate required by the consumer-worker to restore the utility level enjoyed in 1946 given the commodity prices and given nonlabor income in year t. Column (iii) shows that in 1967 a wage rate of almost 1.4 times that received in 1946 was required to restore the utility level in 1946. In fact, as given in column (iv), the actual wage rate in 1967 was 2.25 times the constant-utility wage rate.

The series in column (v) presents that wage rate in year t (w_t^{**}) which restores 1946's utility when consumer goods' prices take their values in year t, but when nonlabor income is fixed at its 1946 level. Column (vii) indicates that a wage double that

*Stanford University. I should like to acknowledge the able research assistance provided by Ray Squitieri and the financial support of the Alfred P. Sloan Foundation.

[1] The entire series for all the years 1929–67 are available from the author upon request.

TABLE 2—POINT ESTIMATES OF PARAMETERS OF THE LINEAR EXPENDITURE SYSTEM AND MEASURES OF PRICE CHANGES IN THE DATA

Group	$\hat{\gamma}_i$	$\hat{B}_i$	Compensated Elasticity	Uncompensated Elasticity	Δp (and Δw)
Durables	.129	.225	−.709	−.934	0.324
Food	.798	.166	−.405	−.571	0.555
Clothing	.547	.061	−.417	−.478	0.392
Other nondurables	.934	.108	−.384	−.492	0.608
Housing services	1.120	.158	−.248	−.406	0.634
Transportation services	.997	.019	−.256	−.274	1.227
Other services	.587	.138	−.410	−.548	1.277
Working hours	2331	−.125	.055	−.070	2.076

Note: Both the compensated own-price (wage) elasticities and the uncompensated own-price (wage) elasticities are evaluated at the sample means Δp (and Ψw) stand for the proportionate change in prices (and wage rates) from 1946 to 1967 as measured by these Department of Commerce data. For further details on estimation method and data, consult Abbott and Ashenfelter.

in 1946 was required to restore 1946's utility. The series in column (viii) shows that the consumer-worker received a wage rate in 1967 some 1.5 times its constant-utility level. This 1.5 increase in real wages is similar to that recorded by two of the *BLS* series given in Table 1, namely, the ratio of average production worker hourly earnings to the consumer price index and the index of real spendable weekly earnings for a production worker with no dependents.

Table 4 reports the results of regressing annual proportionate changes in these constant-utility index numbers of real wages (that is, charges in w_t/w_t^* and in w_t/w_t^{**}) on annual proportionate changes in the published *BLS* series on real wages (that is, those given in columns (i), (iii), and (iv) of Table 1) for the years 1947 to 1967. For the constant-utility real wage index w_t/w_t^*, the standard errors of estimate of the regression equations (which constitute a lower bound on the standard errors of forecast) are approximately 3.5 percentage points, which is the same as the observed standard deviation of $\Delta(w_t/w_t^*)$ over this period. By contrast, the *BLS* wage series

TABLE 3—CONSTANT-UTILITY INDEX NUMBERS OF REAL WAGE RATES

Year	w_t^* (i)	$\sqrt{\text{var}(w_t^*)}$ (ii)	w_t^*/w_o (iii)	w_t/w_t^* (iv)	w_t^{**} (v)	$\sqrt{\text{var}(w_t^{**})}$ (vi)	w_t^{**}/w_o (vii)	w_t/w_t^{**} (viii)
1934	0.460	.0123	0.649	0.697	0.304	.0427	0.429	1.055
1938	0.479	.0120	0.676	0.756	0.345	.0376	0.487	1.050
1946	0.709		1.000	1.000	0.709	-	1.000	1.000
1950	0.796	.0039	1.123	1.221	0.899	.0291	1.268	1.081
1955	0 932	.0329	1.315	1.382	1.065	.0725	1.502	1.209
1960	0.934	.0142	1.317	1.721	1.227	.0869	1.731	1.310
1965	0.888	.0240	1.252	2.235	1.328	.0863	1.873	1.495
1967	0.968	.0204	1.365	2.253	1.424	.1209	2.008	1.532

Note The wage rates are measured in dollars per hour at work. Columns (ii) and (vi) are estimates of the asymptotic standard errors of the constant-utility real wages.

TABLE 4—REGRESSIONS OF CHANGES IN TRUE REAL WAGES ON *BLS* INDICES OF REAL WAGES, 1947–67

Equation Number	Right-Hand Variable(s)	Left-Hand Variable	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\delta}_2$	R^2	*SEE*	*DW*
(2a)	1. Proportionate change in the *BLS* index of real spendable earnings for a worker with no dependents	$\Delta(w_t/w_t^*)$	.023 (.009)	.950 (.347)			.273	.032	2.13
(3a)		$\Delta(w_t/w_t^{**})$	.006 (.003)	.851 (.132)			.675	.012	1.91
(2b)	1. Proportionate change in the *BLS* index of real spendable earnings for a worker with 3 dependents	$\Delta(w_t/w_t^*)$	.029 (.009)	.683 (.392)			.132	.034	1.91
(3b)		$\Delta(w_t/w_t^{**})$	.008 (.003)	.903 (.129)			.710	.011	2.32
(2c)	1. Proportionate change in the ratio of average hourly earnings to the consumer price index	$\Delta(w_t/w_t^*)$	.022 (.013)	.775 (.493)			.110	.035	1.77
(3c)		$\Delta(w_t/w_t^{**})$	−.003 (.004)	1.081 (.174)			.659	.012	2.49
(2d)	1. Proportionate change in production worker average hourly earnings, and 2. Proportionate change in consumer price index	$\Delta(w_t/w_t^*)$	.034 (.021)		.439 (.651)	−.656 (.515)	.129	.035	1.86
(3d)		$\Delta(w_t/w_t^{**})$	.006 (.007)		.846 (.214)	−1.011 (.169)	.709	.012	2.59

Note: Estimated standard errors are shown in parentheses beneath the estimated regression coefficients. *SEE* is the standard error of estimate for the regression equation and *DW* is the Durbin-Watson statistic. $\Delta(w_t/w_t^*)$ possesses a mean of .038 and a standard deviation of .035; the corresponding figures for $\Delta(w_t/w_t^{**})$ are .020 and .020, respectively.

predict movements in $\Delta(w_t/w_t^{**})$ much more closely: for instance, on average, since $\hat{\alpha} \approx 0$ and $\hat{\beta} \approx 1$, a proportionate change in the ratio of average hourly earnings to the consumer price index is associated with the same change in $\Delta(w_t/w_t^{**})$ during these years.

Finally, Table 5 presents estimates of the constant-utility nonlabor income and constant-utility full income. The latter series was not presented in the previous paper so some explanation is in order. In this Stone-Geary context, the γ_h parameter has a natural interpretation as the maximum feasible number of working hours in which case constant-utility full income F^*

TABLE 5—CONSTANT-UTILITY NONLABOR INCOME AND FULL INCOME

Year	y_t^* (i)	y_t^*/y_0 (ii)	y_t/y_t^* (iii)	F_t^* (iv)	F_t^*/F_0 (v)	F_t/F_t^* (vi)
1934	1451.5	1.683	.366	2523.5	1.003	.507
1938	1528.1	1.772	.379	2644.1	1.051	.539
1946	862.2	1.000	1.000	2514.9	1.000	1.000
1950	1096.8	1.272	1.000	2951.8	1.174	1.139
1955	1163.9	1.350	1.012	3336.1	1.327	1.253
1960	1507.8	1.749	1.010	3685.2	1.465	1.430
1965	1821.9	2.113	1.009	3892.1	1.548	1.661
1967	1858.0	2.155	0.991	4115.1	1.636	1.683

Note: Nonlabor income and full income are measured in annual dollars per employee.

may be measured as the value of all endowments required to attain the base period's utility level:

$$F_t^* = (\gamma_h w_t + y_t)^* = \sum_i \gamma_i p_{it} + \left(y_o + w_o\gamma_h - \sum_i \gamma_i p_{oi}\right)\prod_i\left(\frac{p_{it}}{p_{oi}}\right)^{B_i}\left(\frac{w_t}{w_o}\right)^{\theta}$$

The series in column (vi) shows that, given the change in commodity prices and in wage rates between 1946 and 1967, full income measured in this way was some 1.7 times that required to attain 1946's level of utility and some readers may be inclined to interpret this as an index of the overall change in the consumer-worker's welfare.

REFERENCES

M. Abbott and O. Ashenfelter, "Labor Supply, Commodity Demand, and the Allocation of Time," *Rev. Econ. Stud.*, Oct. 1976, *43*, 389–411.

J. H. Pencavel, "Constant-Utility Index Numbers of Real Wages," *Amer. Econ. Rev.*, Mar. 1977, *67*, 91–100.

U.S. Bureau of Labor Statistics, *Employment and Earnings*, various issues.

U.S. Department of Commerce, *The National Income and Product Accounts of the United States 1929–1965*, Washington 1966.

NOTES

The Secretary of the American Economic Association wishes to announce the results of the mail balloting for officers which took place during the fall of 1978. The following individuals took office January 1, 1979: President-Elect Moses Abramovitz: Vice Presidents Irma Adelman and Jack Hirshleifer; Executive Committee members Henry Aaron and Zvi Griliches.

A *Committee to Consider Publication Practices of Book Publishers* has been formed in order to provide the members of the American Economic Association with information on typical issues in negotiations concerning the publication of scholarly work. The committee consists of Professors Martin Shubik, Yale University, Chairman; William J. Baumol, Princeton and New York Universities; and Leo J. Raskind, University of Minnesota, Counsel of the Association. Inquiries or information should be address to Martin Shubik, Cowles Foundation, Yale University, 30 Hillhouse Avenue, New Haven, CT 06520.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1: for meetings scheduled between November and February, July 1· for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting. Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

American scholars and other professionals interested in university lecturing or research abroad (Fulbright-Hays awards) are invited to register with the *CIES*. Registrants receive announcement of the competition in March or April, for appointments to begin twelve to eighteen months later. The general competition for Australia, New Zealand, and Latin American countries closes June 1, and for other countries, July 1. The general composition of the program involving more than 70 countries is expected to be similar to that of recent years. Registration for personal copies of the announcement is now open; forms are available from the Council for International Exchange of Scholars, Suite 300, Eleven Dupont Circle, Washington, D.C 20036.

Beginning in September 1979, the University of Miami School of Law will offer a coordinated program of courses and seminars specializing in law and economics as part of the regular J. D. degree program Courses in the program, developed by the Law and Economics Center, are taught jointly by a professor of law and a professor of economics and have been designed to help prepare law students to deal effectively with the increasing numbers of issues in which economics aids legal analysis. Individuals interested in more detailed information should write to the Coordinator, Law and Economics J. D. Specialization, Law and Economics Center, University of Miami School of Law, P. O. Box 248000, Coral Gables, Florida 33124.

The Interuniversity Consortium for Political and Social Research is initiating a monograph series on social science methodology. The aim of the series is timely publication of innovative work, the scope of which exceeds that of journal article length Each monograph will be between 60 and 125 pages in length The series will be interdisciplinary in scope and will emphasize quantitative research methodology including areas such as conceptualization and design, analysis, formalization, and computer utilization The editorial board consists of Michael Hannan (editor, Stanford University), Christopher H. Achen (University of California, Berkeley), Lutz Erbring (University of Michigan), Robert M Hauser (University of Wisconsin-Madison), Paul W Holland (Educational Testing Service), John E. Jackson (University of Pennsylvania), Karl G. Jöreskog (University of Uppsala), Samuel H. Preston (Population Division, United Nations), Richard Rubinson (Johns Hopkins University), W. Phillips Shively (University of Minnesota), Aage B. Sørensen (University of Oslo), and Nancy Brandon Tuma (Stanford University). Manuscripts (3 copies) should be sent to Professor Michael Hannan, Department of Sociology, Stanford

University, Stanford, CA 94305. Information on subscriptions, and individual and bulk orders, may be obtained from *ICPSR*, Box 1248, Ann Arbor, MI 48106.

The next European Meeting of the Econometric Society will be held in Athens, Greece, September 4-7, 1979 Proposals for papers to be presented at the meeting are being actively solicited by the Program Committee. Those wishing to present papers should submit proposals directly to the corresponding Program Chairman: *Econometrics*: Angus Deaton, Department of Economics, University of Bristol, 40, Berkeley Square, Bristol BS8 1HY, England; *Economics*. Guy Laroque, Unité de recherche, INSEE, 18, Boulevard Adolphe Pinard, F-75675 Paris Cedex 14 Papers should be submitted twofold and be accompanied by three copies of an abstract. Final decisions will only be based on *full* manuscripts. No submission can be considered after April 16, 1979.

Registration forms may be obtained by writing to Emmanuel Drandakis, Rector of the Athens School of Economics and Business Science, 76, Potision Street, Athens 104, Greece. The Society will provide its European members with the registration forms in time

---

The third (December) issue of each volume of the *Review of Social Economy* is devoted to a special theme The December 1979 issue will be "Marx as a Social Economist." You are cordially invited to submit papers. Send all manuscripts to the Editor, *RSE*, DePaul University, 25 E. Jackson Blvd, Chicago, IL 60604 The deadline for submissions is April 15, 1979 A fee of $18 is required to partially cover the cost of processing, as a privilege of membership, members of the ASE are entitled to the submission of one manuscript a year without charge.

The *Journal of Consumer Research* and the Association for Consumer Research announce the annual competition for the best article-length manuscript on consumer behavior based on a doctoral dissertation for which a degree was awarded after July 31, 1976. Anyone interested should write to *Journal of Consumer Research*, University of Illinois at Chicago Circle, P. O. Box 6905, 2152 BSB, Chicago, IL 60680.

The *Journal of Consumer Research* invites prospective authors to submit abstracts for a special issue on the consumption of time, scheduled for publication in late 1980. Material for this issue might deal with any of the following: A review of what is presently known about the consumption of time by different types of consumer groups; Articles on the role of time in the study of consumer behavior; Theoretical models of the consumption of time, preferably with empirical tests of these models; Empirical studies involving hypotheses dealing with the consumption of time; Articles on the effects of use of time on different aspects of consumer behavior, hopefully with empirical support; Articles dealing with approaches to and methods for the measurement of the use of time. Interested authors should send two copies of a 300–500 word abstract, plus two copies of a topical outline, by March 31, 1979 to the Editor, *Journal of Consumer Research*, University of Illinois at Chicago Circle, P. O. Box 6905, Chicago, IL 60680.

A new Center for Quantitative and Comparative Economics has been established at l'Ecole des Hautes Etudies en Sciences Sociales in Paris. Initially, the main themes of research will be aspects of macroeconomic disequilibrium and growth in capitalist and socialist economies, effects of the international economic system on the dynamic paths followed by these economies, and the behavior of government policymakers and planners. International collaboration should be especially fruitful and will be welcomed. Further information is available from Mme. Claire Sarasin, Secrétaire du Centre d'Economie Quantitative et Comparative, 54 blvd. Raspail, 75006 Paris, or (at the same address) from G. de Ménil, S. Kolm, R. Portes, J. Mairesse, and Ch. Sautter.

As part of an effort to establish an information bank and to evaluate computer programs for processing complex statistical data sets, Professors Ivor Francis and J. Sedransk wish to compile a list of developers of statistical packages or programs as well as experienced users of these programs. We are interested in learning about any program which performs one or more of the following tasks in statistical analysis: File building or data management; Editing—error detection, correction, and imputation; Data description and plotting; Estimation of finite population parameters and associated variances for complex sample surveys, Statistical analysis and model building (including subroutine libraries). Please send information to Professor Ivor Francis, Scientific Secretary, International Association of Statistical Computing, 358 Ives Hall, Cornell University, Ithaca, NY 14853.

For the coming academic year, 1979–80, the Japan-United States Friendship Commission will provide a special program of support for a limited number of highly qualified graduate students specializing in the Japanese economy, and who are (1) presently enrolled in or expect to be admitted to a doctoral program in economics; (2) competent in Japanese or willing to be-

come engaged in a program of study to achieve this competence; and (3) American citizens or permanent residents. For further details, write Professor Gary Saxonhouse, Association of Asian Studies, 1 Lane Hall, University of Michigan, Ann Arbor, MI 48109.

Transit Management Program: Postdoctoral research fellowships will be available July 1, 1979, with interest in interdisciplinary study of transportation-related issues and problems. Emphasis is on transit management, planning, operation and maintenance of systems, and energy and environmental relations. Send inquiries to Dr. G. J. Fielding, School of Social Sciences, University of California, Irvine, CA 92717.

The National Institute on Alcohol Abuse and Alcoholism has awarded Rutgers University a grant for an *Alcohol Studies Graduate Program* for the years 1979-81. The program is based at the Rutgers Center of Alcohol Studies, and will enjoy the cooperation of a number of university departments, the Rutgers Medical School, and state and private agencies throughout New Jersey. Applications are solicited for three postdoctoral and two predoctoral fellowships each year. Send inquiries to Dr. Mark E. Lender, Center of Alcohol Studies, Busch Campus, Rutgers University, New Brunswick, NJ 08903. (Telephone 201+932-3510)

Deaths

Marion O'Kellie McKay, professor emeritus, University of Pittsburgh; Falls Church, VA, June 11, 1978.

Norman Townshend-Zellner, California State University-Fullerton, professor, director of the CSUF Center for Economic Education, and Statewide Director of Economic Education, Aug. 30, 1978.

John Zerbinis, chairman, department of economics, University of Ottawa, Sept. 29, 1978.

Visiting Foreign Scholar

Ljubiša S. Adamovic, University of Beograd, Yugoslavia: visiting professor, department of economics, Lehigh University, Jan. 1979-May 1979.

Promotions

M. Akbar Akhtar: chief, Industrial Economics Division, Federal Reserve Bank of New York, Aug. 17, 1978.

Timothy M. Bates: associate professor, department of economics, University of Vermont, Sept. 1, 1978.

Herman A. Berliner: associate professor, department of economics, Hofstra University, Sept. 1, 1978.

Hope C. Corman: assistant professor, department of economics, Rutgers College, July 1, 1977.

T. Windsor Fields: assistant professor, department of economics, Miami University (Ohio), Aug. 23, 1978.

J. Fred Giertz: professor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Gary A. Gigliotti: assistant professor, department of economics, Rutgers College, July 1, 1978.

Robert J. Gitter: assistant professor of economics, Ohio Wesleyan University, Sept. 7, 1978.

Victor P. Goldberg: professor of economics, University of California-Davis, July 1, 1978.

Stephen B. Jarrell: assistant professor, department of economics, Western Kentucky University, Aug 16, 1978.

Patricia H. Kuwayama: research officer and senior economist, Federal Reserve Bank of New York, Sept. 21, 1978.

Thomas K. McCraw: professor of business administration, Harvard Business School, Sept. 18, 1978.

Karl D. Meilke: associate professor of agricultural economics, University of Guelph, July 1, 1978.

R. Allen Moran: associate professor of economics, Lehigh University, July 1, 1977.

Walter Nicholson: professor, department of economics, Amherst College, July 1, 1978.

Nicholas R. Noble: assistant professor of economics, Miami University (Ohio), Aug. 23, 1978.

Ernest H. Oksanen: professor of economics, McMaster University, July 1978.

Edward J. Powers: professor of economics, Northern Michigan University, Aug. 22, 1978.

John Roemer: associate professor of economics, University of California-Davis, July 1, 1978.

W. Earl Sasser, Jr.: professor of business administration, Harvard Business School, Sept. 18, 1978.

Jeannine Swift: associate professor, department of economics, Hofstra University, Sept. 1, 1978.

Marcel Tenenbaum: associate professor, department of economics, Hofstra University, Sept. 1, 1978.

John A. Wenninger: chief, Monetary Analysis Division, Federal Reserve Bank of New York, Oct. 12, 1978.

Betsy B. White: chief, International Reports Division, Federal Reserve Bank of New York, Aug. 17, 1978.

D. Daryl Wyckoff: professor of business administration, Harvard Business School, Sept. 18, 1978.

Administrative Appointments

Lewis J. Altfest: director of research, Lord, Abbett & Co., New York, Oct. 1978.

Alexander S. Balinky: chairman, department of economics, Rutgers College, Sept. 1, 1978.

A. K. Barakeh: chairman, department of economics, University of South Alabama, Aug. 1, 1978.

Janet G. Chapman: chairman, department of economics, University of Pittsburgh, Sept. 1, 1978.

Ward S. Curran: chairman, department of economics, Trinity College (Hartford), July 1, 1978.

Charles A. Lave: chairman of the faculty, School of Social Sciences, University of California-Irvine, Sept. 1978.

Eli Schwartz: chairman, department of economics, Lehigh University, July 1, 1978.

John C. Wassom: head, department of economics, Western Kentucky University, Aug. 16, 1978.

Appointments

Lawrence Abrams, Jr.: economist, Industrial Economics Division, Federal Reserve Bank of New York, Sept. 13, 1978.

S Basheer Ahmed: professor, department of economics, Western Kentucky University, Aug. 16, 1978

James M. Anastos: instructor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Melvin V. Borland, Westminster College: assistant professor, department of economics, Western Kentucky University, Aug. 16, 1978.

Gordon L. Brady: fellow, economics studies program, Brookings Institution, July 1, 1978.

Charles H. Breeden: assistant professor of economics, California State University-Hayward, Sept. 1978.

Peter H. Calkins: assistant professor, department of economics, Iowa State University, Aug. 1, 1978.

John C. Cassidy: professional staff, Hudson Institute, July 1978.

John T Cuddington, University of Wisconsin: assistant professor of economics, Stanford University, Sept. 1, 1978.

Donald J. Cymrot: assistant professor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Mark H. Dalzell: economist, Developing Economies Division, Federal Reserve Bank of New York, Sept. 6, 1978

William A. Dellalfar: economist, Balance of Payments Division, Federal Reserve Bank of New York, Sept. 6, 1978.

Patricia Dinneen, Massachusetts Institute of Technology: associate economist, economics department, The Rand Corporation, Dec. 1978.

J. Colin Dodds: associate visiting professor, department, McMaster University, Aug. 1978.

O Homer Erekson: instructor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Judith C. Fernandez, University of California-Berkeley: associate economist, economics department, The Rand Corporation, Sept. 1978.

Richard L. Fernandez, University of California-Berkeley: associate economist, The Rand Corporation, Sept. 1978.

Frank Giarratani: assistant professor, department of economics, University of Pittsburgh, Jan. 1, 1979.

Amihai Glazer: assistant professor of economics, University of California-Irvine, Sept. 1978.

Gerald A. Gunderson, North Carolina State University. professor of economics, Trinity College (Hartford), Sept. 1, 1978.

L Jay Helms, Massachusetts Institute of Technology: assistant professor of economics, University of California-Davis, July 1978.

Mark M. Hopkins, Harvard University: associate economist, economics department, The Rand Corporation, Feb. 1979.

Jack Johnston: professor of economics, University of California-Irvine, Sept. 1978.

Mark R. Killingsworth: assistant professor, department of economics, Rutgers College, Jan. 1979.

Kenneth Kendall: department of economics and commerce, Simon Fraser University, Sept. 1, 1978.

Laura Kozlowski: professional staff, Hudson Institute, Aug. 1978.

John McCallum: department of economics and commerce, Simon Fraser University, Sept. 1, 1978.

Paul B. Manchester, Catholic University of America: staff economist, Joint Economic Committee, Congress of the United States, July 15, 1978.

Thomas E. Merz: instructor, department of economics, Miami University (Ohio), Aug. 23, 1978.

David A. Moser: instructor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Michael P. Murray, University of California and University of Virginia: associate professor of policy sciences, Duke University.

Frank J. P. Pinton, University of Pennsylvania: consultant, combined energy staff, O.E.C.D./International Energy Agency, Paris, Nov. 1, 1978.

Warren Richmond: economist, Banking Studies Division, Federal Reserve Bank of New York, Sept. 27, 1978.

David L. Reinders: research associate, department of economics, Iowa State University, July 1, 1978.

Karen P. Russell: instructor, department of economics, Western Kentucky University, Aug. 16, 1978.

Christopher A. Sarlo: assistant professor, department of economics, McMaster University, July 1978.

Daniel A. Seiver: associate professor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Dorothy Sobel: economist, Balance of Payments Division, Federal Reserve Bank of New York, Oct. 4, 1978.

Howard D. Soben: lecturer, department of economics, Rutgers College, Sept. 1, 1978.

Daniel A. Seiver: associate professor, department of economics, Miami University, Aug. 23, 1978.

Kevin C. Sontheimer: associate professor, department of economics, University of Pittsburgh, Jan. 1, 1979.

Beverly A. Spikes: department of economics and commerce, Simon Fraser University, Sept. 1, 1978.

John Richards: department of economics and commerce, Simon Fraser University, Sept. 1, 1978.

Stephen E. Usher. economist, Industrial Economies Division, Federal Reserve Bank of New York, Aug. 14, 1978.

Jeanne Wendel: assistant professor, department of economics, Miami University (Ohio), Aug. 23, 1978.

Michelle J. White, University of Pennsylvania: associate professor of economics, Graduate School of Business Administration, New York University, Sept. 1978.

Mark A. Willis: economist, Business Conditions Division, Federal Reserve Bank of New York, Sept. 11, 1978.

Geoffrey Woglom, Boston College: associate profes-

sor, department of economics, Amherst College, July 1, 1978.

Nancy Worth: economist, Developing Economies Division, Federal Reserve Bank of New York, Aug. 23, 1978.

Peter Zadrozny, Federal Reserve Board: assistant professor of economics, Graduate School of Business Administration, New York University, Sept. 1978.

Leaves for Special Appointments

Victor P. Goldberg, University of California-Davis. research fellow, Institute of Advanced Study, Princeton University, 1978 79.

Patric Hendershott, Purdue University: visiting professor of economics, Stanford University, Aug. 1978.

George Horwich, Purdue University: senior economist, Department of Energy, Aug. 1978.

Akira Takayama, Purdue University: visiting professor of economics, Texas A&M University, Aug. 1978.

Richard Weisskoff, Iowa State University: fellow, Social Science Research Council, Sept. 1, 1978–Aug. 1, 1979.

Resignations

David F. Bramhall, University of Pittsburgh, Apr 30, 1978.

Daniel Granot, Simon Fraser University, July 1, 1978.

Arthur Guthrie, Simon Fraser University, June 31, 1979.

Jeff A. Schnepper, Rutgers College, June 30, 1978.

David Shapiro, Ohio State University: National University of Zaire, Sept. 1978.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W), her present place of employment or enrollment her new title (if any), and the date at which the change will occur.

C. Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15, *December*, July 15.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Introduction–Program Chairman

It was George Borts's idea that I should write a brief preface to this issue of the *Papers and Proceedings*, to explain to the membership of the Association how and why they found this particular program in Chicago, and not some other. Thus are precedents created.

Some of my predecessors as President-elect seem to have had the singleness of mind and the strength of character to build their programs around a theme. My approach was rather different. It is best illustrated by the story of a figure in the French Revolution who was sitting at a dinner party in Paris when he heard a crowd go singing and marching by outside. He leaped to his feet and hastily excused himself to his hostess, announcing: "I must follow that mob; I am their leader." I have been around long enough to have the feeling that most of my colleagues are perfectly capable of writing the paper they want to write under any rubric that I might give them. So mostly I asked myself: What are the questions that actually exercise the profession these days? The program that some of you attended in Chicago and that others will read below represents my answer.

Obviously a different President-elect, setting out to do much the same thing, would have ended up with a slightly different program. Some of my prejudices and judgments found their way into these pages, at least through my choice of organizers of the individual sessions. Of course one appeals to one's friends, but that is hardly a random sample either. My view is that the main business of economics is the use of theoretical principles to throw light on particular policy problems or particular empirical puzzles. Some of my best friends are theorem provers and institution watchers, but I am trying to be candid about my own prejudices. That one is reflected in the largest number of the sessions reported here, particularly the ones devoted to fiscal and monetary policy, regulation, energy, inflation, the demand for money, Social Security, employment policy, ocean policy (I am an amateur lobsterman), and What Economists Think (a revolutionary concept), and especially in the stroke of luck that put Fred Kahn's name into my head as a potential Ely Lecturer.

The next largest group of sessions is more frankly theoretical: the ones on the non-market-clearing paradigm (which came out a bit like the wolves appraising Little Red Riding Hood, but virtue will triumph), wages and employment, the economics of information, and the theory of industrial organization. It seemed to me that the convention is a good place for the rest of us to see what the Bright Young Things are doing, and where the newer theoretical principles of the future are likely to come from. I have to admit that if I didn't think some of these ideas would later prove Relevant I would have been less likely to seek out sessions.

There are two trade-union-type sessions, the regular one on economic education and another on the academic labor market for economists. We are a professional association (in addition to being a fount of wisdom) and those seemed interesting and appropriate to me. I hope some successor will do a follow-up for nonacademic economists.

The remaining sessions were targets of opportunity; almost by definition they represent what at least some segment of the profession found live and interesting.

It may be that this volume of the *Papers and Proceedings* will be unusually widely read, since the members did not exactly cluster around Chicago in August like moths round a flame. I hope it will be found interesting. If not, don't blame me; I was just following the mob.

ROBERT M. SOLOW

Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninety-first annual meeting of the American Economic Association. The *Proceedings* consist of the record of the business activities of the Association in 1978: the annual membership meeting; the March and August meetings of the Executive Committee; and reports of various Association officers and committees. As with the Notes section in each issue of the *American Economic Review*, they are published to keep the members informed and encourage them to participate in the Association's affairs.

The *Papers* constitute the greater part of this volume. They are roughly equivalent to two regular issues of the *American Economic Review*, but are published under different procedures. About a year in advance, the Association's President-elect (in 1978 Robert Solow, in 1979 Moses Abramovitz) acting as program chairman, decides on the topics of sessions at which papers will be presented. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. The program chairman also sets limits on the length of papers at various sessions, and invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the topic of the session, and asks others to give comments on the papers. Some of the sessions are devoted to contributed rather than invited papers. The program chairman decides at the time of organization which sessions are to be printed. This year the program chairman decided to devote the issue entirely to the invited papers, to the exclusion of comments and contributed papers.

The papers to be published are sent to the editorial office of the *Review*; the editors of the *Papers and Proceedings* check them for length and content, and send the authors comments and suggestions.

The rules under which these papers are published are quite different from those governing articles appearing in regular issues of the *Review*. Their length is strictly controlled. Except in unusual circumstances they must be less than 4,000 (and sometimes even 3,000 or 2,000) words in length. Their content and range of subject matter reflects the wishes of the program chairman to explore and expose the current state of economic research and thinking. In many cases they are exploratory and discursive rather than definitive presentations of research findings. While we do edit the papers to improve content and style, to satisfy space constraints, and to eliminate repetition, we do not subject the papers to any refereeing process, and publication of any paper received prior to the printing deadline that satisfies space requirements is virtually guaranteed.

We would refuse to publish a paper if we concluded after reading it that it was utterly without merit; no paper has yet been rejected on these grounds. The Executive Committee has established another ground for rejection: if a paper cannot be cut to meet space requirements, we may ask the author to authorize its consideration for publication in a regular issue of the *Review* (subject to the usual refereeing process). Or the author may be asked to withdraw the paper and submit it elsewhere.

These policies serve a number of important purposes: the papers can be published without the long delays imposed by the refereeing process. They are short papers, covering a wide variety of subjects, and in most cases can be understood by nonspecialists. Authors receive a chance to report on research just completed, discuss topical subjects in an informal way, and to summarize longer forthcoming publications. Readers get a chance to browse among a large number of articles which are outside their major areas of interest, but which are not as specialized or as technical as those sometimes found in the regular journals. And while the papers are not refereed, they do provide an accurate picture of the state of thinking in many of the fields of economics.

GEORGE H. BORTS
JAMES A. HANSON

RICHARD T. ELY LECTURE

Applications of Economics to an Imperfect World

By ALFRED E. KAHN*

Coming to me as it did after almost a decade's absence from the academic profession, I accepted this invitation only with trepidation. While the profession has been extending the frontiers of economics, I have been operating deep within its margin, first discovering how dismal our science really can be as it applies to the finances of private universities, and then, during the past four years, applying to the real world economic principles that Alfred Marshall would have had no difficulty recognizing.

I have no particular interest in describing the first of these experiences. Its only lessons were that the laws of economics are truly made of iron; and that any organization that hopes to make the best use of its limited resources had better be organized more hierarchically than a university. The experience of being a practitioner of regulation, in contrast, has been immensely satisfying, because it has afforded almost unlimited opportunity for the application of simple micro-economic principles to the real world.

The applicable principles are easy to characterize: that economic efficiency calls for prices equated to marginal social opportunity costs; and that, whenever it is technologically feasible, competition is the best institutional mechanism for achieving that result, as well as for minimizing X-inefficiency and ensuring the optimum rate of innovation. What has been especially intriguing about my experience is that it has embraced two quite different regulatory situations—one, the traditional public utilities, where competition seems for the most part infeasible, and the economist-regulator is moved to play an active role in trying to produce efficient results; the other, airlines, in which it appears the prime obstacle to efficiency has been regulation itself, and the most creative thing a regulator can do is remove his (and her) body from the market entryway.

But the process of applying these principles—even of simply getting out of the way—has been far from simple. The slate on which the economist-regulator writes is scribbled with the scratchings of lawyers, jurists, and politicians; the world to which he would apply his principles is excruciatingly imperfect and resistant; and the compass he needs is one that would help him thread his way through the thickets of second best. The really challenging job is deciding not what the ultimate economically rational equilibrium should look like, but what is economically rational in an irrational world, and how best to get from here to there. That, too, turns out to be a kind of frontier; and life on it is full of excitement.

I. Problems in Regulating Monopoly

A. *Problem 1: The Institution of Regulated Monopoly Itself*

Here in Chicago, it would be supererogatory for me to linger long over the defects of the institution of regulated monopoly: the sufficient summary, which traces back at least to Henry Simons, is that it combines the worst of both worlds—the evils of monopoly with the stultification of the profit motive. I would add, with George Stigler and Richard Posner, that it offers an irresistible opportunity to use price—typically very imprecisely and inefficiently—as an instrument for the redistribution of income.

*Chairman, Civil Aeronautics Board. I acknowledge gratefully the criticisms and suggestions of Elizabeth E. Bailey, Paul L. Joskow, James C. Miller III, and Irwin M. Stelzer.

One of the most sobering lessons of my experience with public utility regulation was the progressive realization that my most energetic initiatives were little more than feeble efforts to compensate for the inherent defects of the institution over which I was presiding.

One of my proudest accomplishments, for example, was the progress we made in requiring the electric and telephone companies in New York to introduce marginal cost-related prices. If you are a large residential user of electricity on Long Island, for example, instead of paying the previous flat charge of so many cents per kilowatt hour, you will soon—if the courts allow—pay rates varying between 2½ cents at night and 30 cents on summer days when the temperature gets above 83°. As a specific example of the encouragement that this kind of pricing will offer to rational choices between consumption and abstinence, energy and insulation, the use of fuels or the sun, consider what the introduction of that marginal cost-based 12 to 1 ratio does to the likelihood of storage cooling being developed and introduced commercially. Again, the business customers of the New York Telephone Company now have to pay for their local calls on a timed basis; they can no longer ignore the fact that additional minutes of conversation have a positive marginal cost. Residential users are offered a similar pricing system, with the inducement of reduced basic charges.

In trying to introduce changes like this we encountered strenuous resistance, not just from large users who thought they would be disadvantaged by them, but from the utility companies themselves. Why would the electric companies cling to a declining block rate structure, without reference to the time of consumption, when it appeared, particularly at times of peak demand, that sales in the ultimate blocks were markedly below marginal cost, and the result was to intensify the financial squeeze to which they were in any event being exposed by the combination of inflation and regulatory lag?

I can think of only two reasons. First, bureaucratic inertia; and second, a lingering assumption that it was in their interest to promote additional sales that require additional investment, for the familiar reasons exposed by Harvey Averch and Leland Johnson, among others. But both of these phenomena are themselves surely the consequence of regulated monopoly—of the absence of competition, and of regulation on a cost-plus basis, with allowable returns reckoned on invested capital. So a plausible case can be made that regulation itself was one of the imperfections we were trying to overcome—that all this furious activity to reform utility rate structures was itself necessitated by regulation.

This same observation applies, I think, to our very strenuous attempts to attack the problem of X-efficiency—our introduction of management efficiency audits; our embodiment of productivity targets in the rates we set; and our efforts to force surprisingly reluctant separate gas and electric companies to engage in more comprehensive integration of their investment and operations. Unregulated monopolists would presumably have strong incentives to hold their costs down and to buy rather than produce for themselves whenever the marginal costs of the former were less than of the latter.[1]

A clear understanding of the limits of what

[1]I cannot refrain from citing one last example, in which I took particular pride, but which illustrates even more clearly the point I am making here. From time to time, we at the New York Public Service Commission found ourselves confronted with requests by small water companies for rate increases in the range of 200 to 300 percent, which, to our astonishment, our staff testified were necessary to enable them to cover their costs and provide a reasonable return on investment. It was very difficult to believe, in these cases, that costs had increased by percentages of that order of magnitude during the period in which the then current rates had been in effect. The explanation was not hard to find While separate legal and accounting entities, the companies in question either were or had been appendages of real estate developers, who got into the water business because most of their customers were unwilling to buy developed lots and houses without an attached water supply. Whatever they earned, they earned not on the water systems as such, but on the combined operation. Now they were proposing to make the water operation compensatory by conventional regulatory standards. It proved fairly simple to explicate the sense of injustice expressed by some of their indignant customers. The price that purchasers had paid for the developed lots or houses must have reflected, explicitly or implicitly, the

regulation can accomplish under monopoly has the very healthy effect of making an economist-regulator anxious to seize every possible opportunity to render it unnecessary. We took major steps in New York, for example, toward opening the market for telephone terminal equipment, including interior wiring, to free competition; this particular part of the industry, we were convinced, could be effectively competitive. I will say quite a few words later about the similar steps we have been taking in the field of air transportation.

B. *Problem 2: Second Best*

Prominent among the opponents of marginal cost pricing of electricity were a group of large industrial and commercial users, some of them out of ignorance and inertia, others understandably fearing it would be used to discriminate against them, and others simply unwilling to pay the cost of servicing them. They hired a number of economists to proclaim solemnly that Richard Lipsey and Kelvin Lancaster were on their side: it would be inefficient, they asserted, to price electricity at marginal cost—which has almost certainly, after so many years of inflation, come to exceed average revenue requirements, as traditionally determined—when the prices of natural gas and oil are both being held below *their* marginal costs.

The observation was, of course, pertinent. My own provisional answer has the following parts:

1. First of all, second best argues no more persuasively against moving prices to marginal cost than it does against leaving them where they are.
2. The field price of natural gas is, indeed, being held below marginal opportunity cost; but since, for that very reason, gas is being physically rationed, pricing electricity up to *its* marginal costs is not likely to produce a substantial diversion of consumption to this underpriced substitute.
3. The price of domestic crude oil, similarly, is clearly being held artificially below the marginal cost to the American economy, which is the delivered price of imports. But the regulation affects only a declining fraction of total domestic supply, which constitutes in turn only a fraction of the retail price.
4. Moreover, oil is a major input in the generation of electricity, and it takes three Btu's of oil to produce one of electricity. This fact, along with the external costs (in terms, for example, of our national terms of trade) of sharply rising oil imports, argues powerfully for pricing electricity at marginal cost, at least where oil-fired generation is marginal.
5. Other less obvious but extremely important substitutes for electricity are all priced at something like their respective marginal costs—insulation, the incorporation of additional efficiency in electric appliances and equipment. The choice among these particular substitutes cannot be made efficiently unless electricity itself is similarly priced.

In short, the presence of governmentally imposed distortions in other parts of the economy does not, as the opponents of marginalism seem to think, render economic prescriptions invalid. They merely make the analysis more difficult.

---

price they were being charged for water, and certain expectations about its future course. It seems a reasonable assumption that the purchasers had no reason to expect their water rates to go up more than costs. If that assumption is correct, the inference is inescapable that to grant a water company associated with a real estate developer a rate increase by more than costs had increased since the time of purchase would have involved permitting a double recovery of the original investment—once in the selling prices of the houses, and the second time, by courtesy of the Public Service Commission, in the price of the water itself. The solution we developed was to require applicants for rate increases to justify them in terms of the *increases* in costs that they had incurred over some reasonable period of time in the recent past. This involved establishing a presumption that when the rate increases justified by the rate base/rate of return criterion exceeded those demonstrated cost increases, the differences were ipso facto evidence of an attempted double recovery—that is, that some portion of the capital dedicated to providing water had already been recovered in the sale prices of the lots and houses. It took some modest ingenuity to protect customers in this way. But it was sobering to reflect that what we were protecting them from was an irrationality that flowed from the traditional method of regulation.

C. *Problem 3: Subsidization*

The same, of course, is true of legislative decisions to subsidize or cross subsidize certain kinds of consumption. These decisions usually leave a determined regulator a considerable margin of discretion in deciding what shall be subsidized, how much, and how. For example, Congress is determined to spend as many as a hundred million a year of taxpayer dollars to provide air transportation service to relatively small and isolated communities, over relatively thinly travelled routes. There is no point in fighting that policy, particularly when some case can be made for it on grounds of the external benefits of linking the country together and avoiding urban congestion. But what the Civil Aeronautics Board (*CAB*) has done is explain to Congress how it may get what it wants more efficiently, first, by permitting free entry of air taxis and commuter airlines—which can often perform these particular services at much lower cost than the certificated carriers; and, second, by specifying the subsidized services we want to purchase and attempting to purchase them at minimum cost, rather than, as under the present system, essentially by making good the revenue deficiencies of the carriers certificated for this purpose (this description does less than justice to the progressive efforts by the Board over the years to refine the methods of subsidy determination, but it will have to suffice).

Similarly, society seems determined to have basic telephone service provided at less than cost and, even worse from the efficiency standpoint, through internal subsidization. The reasons, when they are articulated at all, are usually stated in terms of externalities (my telephone is valuable to me only as it enables me to reach others) or "social welfare." A regulatory commission can be persuaded, however, that these cases for subsidization apply validly only to the opportunity to *receive* unlimited numbers of calls, and possibly to place some minimum of outgoing ones; but that they provide very little justification for subsidizing what passes for basic service in most places in the country—which typically includes the opportunity to place an unlimited number of local calls, of unlimited duration, at no extra charge. Confining the subsidy to the former, truly basic service alone, while introducing individual charges for each additional local call and for additional minutes of calling, minimizes the inefficiency that results from holding rates below marginal costs, and has the additional satisfying effect of rewarding with lower bills people who are willing to exercise some restraint in the costs they impose on the system.

Economic logic can also be very fruitfully applied to devising a least distorting method of financing this internal subsidization. The traditional method has been by charging markedly above marginal costs for interstate calls, on the ground, among others, that since the very costly installation at the subscriber's end is used for both intrastate and interstate calls, it is only "fair" that both share the responsibility for covering its costs.[2] The consequence is that every time a telephone or a switchboard is installed, some 20 percent of the capital cost is automatically transferred to the interstate revenue requirement, there to be imposed upon long-distance calling.

I can tell you from experience it is possible to persuade regulatory commissioners that it is inefficient to levy the cost associated with these installations on *usage* of any kind—whether interstate or intrastate. The distortion is particularly inefficient in the case of the telephone, because it seems clear the marginal costs of long-distance communications are far below average revenue requirements. And, the Bell System pointed out, this transfer inflated interstate toll charges in 1974 by 40 percent! Since the entire cost is incurred at the time of installation, and the marginal cost of *using* the equipment thereafter is zero, we in New York State transferred hundreds of millions of dollars of these annual revenue requirements to the monthly lump sum charge.[3]

[2]The most obviously "fair" basis for doing so, so the argument has run, is (a) in proportion to the relative use of the equipment for these two purposes, and (b), since a minute of interstate use is more "valuable" than of local, by factoring the former minutes up by some multiple—which has itself been increased over the years.

[3]I must not exaggerate our achievement: since no state can afford to pass up the subsidy from interstate, we were

II. Problems of Managing a Transition to Competition

During the last fifteen months, I have been coping with a somewhat different set of problems—those posed by the transition of the airline industry from a regime of rigid governmental protectionism and cartelization to one of free competition. I have very little to add to the extensive literature endorsing that goal. It provides very little guidance, however, for getting there—specifically, for coping with the inevitable distortions of a transition that is going to take some time, partly because the law under which we operate still requires us to find, case by case, whether granting each application for entry accords with the "public convenience and necessity," while giving each incumbent competitor—exercising procedural rights that trace back at least to the Magna Carta—an opportunity to argue that it will not.

What I propose to explain here is my conversion from a belief that gradualism is actually desirable to an advocacy of achieving as quickly as possible something as close to total deregulation as the law will permit.

My original attitude was based, first, on simple intellectual caution; second, on a desire not to discredit deregulation by showing an insensitivity to the fears of both Congress and the financial community about what a sudden total immersion in the waters of competition might do to the financial health of the industry, especially since it had just emerged from five or six years of dismal earnings. Finally, I thought that, since the airline companies had lived in a protectionist hothouse for forty years, their managements had to have time to plan for the new competitive era—to rationalize their operations, to meet the additional competition to which they would become subject, and to be ready to grasp the competitive opportunities that would shortly be presented to them.

I was not unaware, even at the outset, of the possible distortions of a gradual process. The theory of second best tells us that if we want to go from point A to point C, it is not necessarily socially efficient to go part way. And I will shortly be presenting several concrete illustrations of the principle. To anticipate the conclusion, however, I originally thought that meant that we ought to move very cautiously, examining the results every step of the way, in hope of minimizing the disruptions and distortions of the transition; my present conviction is that it means we must make the act of faith and move just as rapidly as possible all the way to C.

A. *Problem 1: Unequal Competitive Abilities*

The airline industry carries over into its present an incredibly complicated burden of restrictions and impediments from the past. The most important explanation of the differences in cost among different carriers is their respective bundles of operating authority and restrictions, and the kinds of routes and route structures they serve—long haul or short, in thick markets or thin. Moreover, the ability of one carrier to compete successfully over a particular route with another will be heavily influenced by the extent to which it and its rivals have available customers from their own feeder routes that they can readily funnel into their own operations, and rights to routes beyond onto which they can feed their passengers, thereby permitting them to fatten up their flight schedules on the contested ones. Continental Airlines, for example, which lacks route authority eastward of Chicago, argues strenuously that it would be a serious competitive disadvantage if carriers with richly diversified feed into O'Hare Airport from the East were free to invade the comparatively few routes to the West that contribute the bulk of its profits.

Route structure is, indeed, the dominant influence on relative unit costs; but carriers

---

not in a position to correct the inefficient inflation of interstate long-distance calling. What we did, however, was transfer that benefit from the decision to install terminal equipment to the charge for basic service; and as we made progress toward charging for local calls on a unit and minute basis, fully reflecting marginal cost, so we were able to concentrate the subsidy increasingly on the one portion of the service that seemed to us, following the logic I have already outlined, the most worthy of subsidization—and also, because the demand for basic service is probably comparatively inelastic, the place where subsidization produces the least inefficiency.

compete over *particular* routes. And while the one with the most feed can flow traffic over particular contested routes, and in this way beef up its schedules to the disadvantage of its rivals, there is ample evidence that it is not the biggest carrier, with the most ample feed and beyond operations, that uniformly enjoys competitive superiority. All three of Continental's competitors between Chicago and Los Angeles, for example, have rich feed from the East; yet Continental competes with them very effectively.

If there are advantages of integration, there are also powerful economies of specialization. A lack of feed and beyond traffic did not prevent Pacific Southwest Airlines from becoming the dominant carrier in the Los Angeles to San Francisco route, or Southwest Airlines from duplicating that success between Dallas and Houston; and, it is interesting to observe, one of Eastern Airlines' most profitable routes is the shuttle between Washington, New York and Boston, in which it has surrendered any possible advantages of single plane service, feeder or beyond operations.

So far as I know there is no objective basis for deciding which of these situations is more likely to prove typical—the one in which size and network economies are decisive, or the one in which the specialized carrier will have clear advantages. Most markets undoubtedly fall in between. In market after market today, carriers of widely varying sizes and degrees of integration meet in head to head competition; there is no systematic evidence that this cannot continue indefinitely. Perhaps the only conclusion one can and need draw is that under a competitive regime, these various kinds of market situations will sift themselves out automatically, with various kinds of suppliers emerging successfully on the basis of their respective advantages and handicaps in each. Our uncertainty about the outcome of the competitive struggle is no reason to prevent its taking place; the only sensible prescription is to give the competitors freedom to slough off their artificial handicaps by entering and leaving markets, as they please.

Moreover, if we cannot *predict* how these offsetting advantages and handicaps of the several carriers are likely to work out under a regime of free entry, it seems to me even less likely that we can hope to achieve the most efficient performance of the transportation function by *prescribing* how the thousands of markets should be served, as the proponents of the status quo would have us do. I find it difficult to see how these uncertainties tilt the balance in the direction of a reliance on predictably ignorant regulation in preference to an uncertainly predictable market process.

B. *Problem 2: Distortions in Moving Piecemeal*

Some carriers profess not to worry about their ability to survive a competitive struggle if we were able to deregulate promptly and totally; but they argue strenuously against our decreeing totally free entry into markets on a case-by-case basis, in the order in which applications happen to be presented to us.[4]

The problems they envisage seem to be of two kinds. First, a Continental or a National argues that the market-by-market approach to free entry may subject it to waves of competition in its particularly important markets, while delaying its entry into lucrative new markets. I see no reason to assume, however, that the order of our proceeding will have a systematic bias of this kind. In fact, our two most dramatic proposals to open large numbers of markets to multiple permissive entry—involving service to and from the underused Chicago Midway and Oakland airports—have been ones in which the great bulk of the traffic will be purely turnaround; in which, therefore, feed and beyond rights will be of little importance; and in which prominent among the applicants are carriers with no such route systems at all.

The second fear is that if only some markets are opened to entry and not others, all the competitive energies of the industry will concentrate on them, resulting in exces-

[4]See, for example, the brief of Continental Airlines ("Improved Authority to Wichita Case") on the issues of delayed multiple awards and/or conditional permissive back-up authority.

sive entry and investment. All this comes down to is the destructive competition scarecrow: there seems to be a general belief among defenders of the present regulatory regime that there is something about airplanes that drives businessmen crazy—that once the *CAB* removes its body from the threshold, they will rush into markets pell-mell, like lemmings, without regard to the size of each, how many sellers it can sustain, and how many others may be entering the same time. It doesn't happen in other industries; there is no reason why it need happen in air transport.

It remains undeniable, however, that the gradual approach, market by market—which may be forced on us by the Federal Aviation Act—must involve distortions. So long as deregulation is incomplete, so long as the certificate of public convenience and necessity continues to have an exclusionary and therefore a market value, some of the airlines assure us, they will apply for more licenses than they can operate economically, and operate under them sufficiently to ensure that they are not taken away; and they will flood markets with more service than is economic in order to preclude competitive operations by others, in the hope of being able in the future to reap the rewards of the monopoly power they achieve and preserve in this way.

The only rational answer to this possibility is to demonstrate convincingly that the value of these franchises is going to be zero. Then there will be no valuable pieces of paper to fight for with uneconomic operations, and no future monopoly gains to offset against the costs of present predation. It is of course necessary to convince the companies that this is going to happen, but the way to do that is to open markets to free entry—and that is what we are doing. Moving as rapidly as possible to a system of universal free entry—and exit—is the way to deal also with the asserted inequality of competitive abilities and opportunities during a slow transition: make the transition rapid; move quickly, on as broad a front as possible, to permit all carriers to slough off the restrictions that limit their operating flexibility, to leave the markets they find it uneconomic to serve, to enter the markets they want to enter. The legal uncertainties are far from negligible; but there is more than a small chance the courts will let us define the "public convenience and necessity" in this intelligent way, provided we explain very clearly to them exactly what we are doing and why. (See my article, "A Paean to Legal Creativity.")

C. *Problem 3: Do Innovators Need Protection?*

Despite legend to the contrary, the *CAB* has, over its forty years, admitted a large number of new domestic airlines into scheduled operations; still, the five we have licensed in the past two months (see "Chicago-Midway Low Fare Route Proceeding," "U.S.-Benelux Exemptions," "Application of World Airways (Guam Exemptions)," and see also, "Applications of Colonial Airlines, Inc.") to compete directly with the trunks and regional carriers, and our adoption both in specific cases and in general principle of the policy of admitting all applicants on a permissive basis clearly reflect a dramatic change in entry policy.[5]

[5]The Board has over the decades admitted to the industry on the order of ninety new airlines by certificate, and several hundred more by exemption. These were, however, almost exclusively, authorizations to provide specialized or geographically limited or otherwise circumscribed services in peripheral markets; they include subsidized small community, charter, all cargo, small aircraft commuter, Alaskan and helicopter service. Only one at the time of its entry (Trans Caribbean Airways) infringed on the principal passenger markets served by the "trunks"—the sixteen or so major carriers who received grandfather licenses in 1938, now reduced to ten through merger. While the attrition rate among these newly admitted carriers has been very high, it is from survivors of this group that the small amount of new entry into markets served by trunk carriers has occurred. Trans Caribbean, for example, which the Board authorized in 1958 to serve the New York-San Juan market, was an exempted "irregular" carrier which converted charterlike operations into regular scheduled services. It has since merged with American. Two Alaskan carriers were authorized to operate in competition with existing trunks on Alaska-Seattle routes. And, most important, today the eight major "local service" carriers serve about 9 percent of the total forty-eight state passenger market, and, over a twenty-year period, the Board has gradually authorized them to compete with trunks in the smaller and shorter haul trunk markets. While, therefore, the

Two of these recent certifications raised the old but still challenging question of the compatibility of pure competition with innovation. These were the extremely attractive, novel applications of Midway Airlines and Midway (Southwest) to provide commuter service between the essentially unused Midway Airport in Chicago and several midwestern cities, at basic fares approximately 50 percent of the level that the *CAB* had theretofore uniformally prescribed. The first of these was a paper company, the second an affiliate of the highly successful Texas intrastate airline that had pioneered in the introduction of the same kind of highly efficient, specialized low-fare commuter-type service as was being proposed here. The two applications were shortly met with filings by other carriers to serve some or all of the same markets, and with declarations by incumbents already licensed to serve Chicago that they would "meet the competition"—i.e., reduce their fares in these markets and in some cases make use of Midway Airport as well—at least one of them before the two new carriers could even hope to obtain certification from the Board, let alone acquire the necessary aircraft.

Several civic parties urged us to protect one or both of the innovators by giving them for a year or two the exclusive rights to serve Midway Airport; some of them originally proposed that we also prohibit incumbent carriers even from matching the low fares at O'Hare. The innovators, they argued, needed and deserved a period of exclusive right to exploit their new idea. If, instead, we were to permit the many larger and better established rivals to emulate—indeed anticipate—them immediately, how could we be sure, once the two upstarts were aborted or eliminated, that the existing carriers would not drift back to O'Hare, as they had done in the past, each of them finding it in its own interest to concentrate its flights on the airport where its passengers would have the greatest possible likelihood of making connections?

Once again, we confronted the distortions inherent in gradual deregulation. Despite our use of extraordinarily expedited procedures, these applications had been pending for almost two years. In effect, therefore, our certification process was acting like a patent system in reverse: whereas under a patent system, the innovators would have been rewarded for the required public disclosure of their plans with a period of exclusive right to exploit them, under the Federal Aviation Act it would be their already certificated rivals who would be given the head start!

Time will not permit even a summary of the reasons that led us finally to reject this plausible argument. The ultimate consideration was that we were not persuaded it was necessary to grant this period of exclusivity in order to ensure the successful commencement of the service. Instead, therefore, we grasped the opportunity to make our first major grant of universal authority to all applicants, in the belief that this would ensure the fullest and most rapid possible exploitation of the market, and that the competitive market would do a better job than we of deciding what service, and how much, would be economically feasible, and which carriers would be the best equipped to provide it.

But this is not the end of the story. Partly because of the very distortions of the transition that led some of us to think long and hard about giving the innovators a head start, we decided to move even faster and farther than we had originally contemplated. Having

frequently encountered statement that the *CAB* has since 1938 permitted no new entry into the airline business, or at least into the portion operated by the so-called trunk carriers, is wrong in detail, it correctly describes the general spirit of the Board's general practice during this entire period. It is important to recognize that the airlines constitute a growth industry almost second to none in American business: from 1938 to 1977, revenue passenger miles (*RPM*) increased almost 300 fold. Yet despite this growth the original grandfather carriers continue to provide over 90 percent of the *RPM* in the forty-eight state domestic market. Just under half of the 9 percent share of the certificated local and regional carriers is in markets never served by the trunks, or previously abandoned by them. To repeat, except for Trans Caribbean, essentially none of the firms admitted were admitted into direct competition with the trunks in the forty-eight states. The entry policies of the Board during the last few months—and especially our moves to admit all applicants on a permissive basis (see our "Oakland Service Investigation")—therefore represent a radical break with the past forty years.

decided upon the policy of multiple permissive entry into all the six markets to which we had narrowed the case, in order to make it manageable, we then tentatively decided to extend our permission to an additional seventeen Midway markets by summary procedures. (See "Chicago Midway Expanded Service Proceedings.") One important consideration was our desire to minimize the undeniable possibility that incumbent carriers would blanket all the available opportunities and so preclude operations by Midway and Midway Southwest. The idea was to open up so many the incumbents would simply run out of blankets. I was therefore enormously gratified with the reaction to this decision by Midway, which had flatly asserted during the case that it could not get off the ground without exclusive authority: "Kenneth Carlson, one of the . . . owners and its marketing vice president, said . . . that by expanding the available routes to 23 from six, the CAB would give Midway Airlines ample market prospects. 'It's going to be harder for (established carriers) to grab us in a bigger fishbowl,' Mr. Carlson said."[6] Precisely as we intended.[7]

D. *Problem 4: Liberalizing Entry when Airport Space is Inefficiently Rationed*

The certification of Colonial Airlines, our third wholly new entrant this year, provides a quite different but even more poignant illustration of the problem of determining what constitutes rationality in an irrational world. Colonial applied for authority to provide commuter service between Morristown Airport, in northern New Jersey, and Washington and Boston. Our Administrative Law Judge concluded the service was needed, but recommended against certification because Washington National Airport is badly congested at peak hours, its slots are allocated by agreement among the certificated carriers, and there was a real danger that a certificated Colonial, carrying at most fifty-six passengers per flight, would be able to claim a slot at the expense of some other carrier carrying several hundred to or from somewhere else—a rational second best kind of calculation. In addition, the city of Newark importuned us to turn down the application on the ground that there is excess capacity at Newark Airport, nearby.

The basic problem is that airports are for the most part separately owned, each of them charges landing fees based on its own embedded costs, and few if any follow peak pricing principles even modestly. So the choices by carriers and passengers of flying times and airports are blithely uninfluenced by what must be vast differences in marginal opportunity costs, except to the extent that rationing by intercarrier agreement produces the same results—which seems extraordinarily unlikely.

We were unwilling to settle for a very poor second best. We certificated Colonial (see "Applications of Colonial Airlines, Inc."); we advised Newark to put pressure on the New York Port Authority, which operates all three metropolitan airports, to introduce marginal cost pricing—which would mean reducing Newark landing fees sharply and increasing them at the other two; we began a reconsideration of the antitrust exemption we had been routinely giving the carriers to get together and allocate airport slots; and we have initiated consultations with the Federal Aviation Administration to explore the possibility of devising schemes—preferably rational pricing—to ensure a more efficient allocation of scarce take off and landing space. In this

[6]See *Wall Street Journal*, July 13, 1978.

[7]This decision of ours in the Midway case, and a similar tentative one proposing open competitive entry between a large number of cities and the similarly underused Oakland Airport, does not mean that we have decided upon a universal policy of free entry. Even apart from the uncertain legality of such an across-the-board policy under a statute that seems to require public convenience and necessity findings market by market, we must confront the arguments we continue to encounter, especially on thinner and more clearly marginal routes, that the only hope for their development lies in confining authority to a single carrier, on the ground that no company would be willing to invest the resources in exploring and developing such markets if, the moment it had demonstrated the possibility of profitable operations, it were subjected to competitive entry by others. It is conceivable that some markets will not be developed unless the benefits of innovating and developmental activity can at least for a time be appropriated exclusively by the innovator.

case first best is surely much better than second.[8]

E. *Problem 5: Calibrating the Liberalization of Pricing and Entry*

If one is not to remove all controls at once, it is important to try to see to it that controls over price do not get removed too rapidly or too slowly relative to controls over entry. The need for this caution is most obvious in removing price ceilings in the continuing presence of monopoly power.

On the down side, I am not certain that the increasingly permissive attitude that the Board has taken during the last year toward price reductions—to the point of almost total *laissez faire*—while new entry by would-be competitors continues to be embroiled in the still maddening slow certification process, has not caused us to miss the opportunity for a restructuring of the industry along more competitive lines. It is possible that by permitting incumbent carriers during the last year to introduce a vast variety of discount fares—many of them highly discriminatory and appealing to the same elastic demand travelers as the charters and Freddie Lakers—we may have enabled them to foreclose entry into the provision of uniformly low-fare scheduled service by the supplemental carriers, some of whom have been seeking this authority for years.

The final returns are not in on whether we have moved too quickly, although I believe we have not. The pertinent observation, in any event, is that the logic of events[9] has driven us in the direction of trying to synchronize the processes of decontrolling price and entry by speeding up the latter rather than moderating the former—in the direction, once again, of speeding up the transition. Equalizing restrictions turns out to be like equalizing the two sides of a mustache: one can do it much more rapidly by cutting down on the longer side than by extending the shorter one!

F. *Problem 6: Maintaining an Efficient Balance of Price and Nonprice Competition*

It would have been equally undesirable to have liberalized entry more rapidly than pricing.

When I came to the Board, it had pending before it over 600 applications for route authority of varying degrees of vitality and sincerity. Only a handful of these involved a direct promise of price competition—a small hand with only a few fingers. The others were simply applications to enter given city-pair markets and offer service in competition with a single or very small number of incumbent carriers at the same prices.

One lesson we have learned from the history of airlines is that in the absence of price competition, rivalry among carriers tends instead to take the form of costly improvements in service—particularly additional scheduling. An increase in the number of carriers in a particular market seems to have been correlated with a decline in load factors—an increase, in other words, in cost-inflating scheduling rivalry—producing an apparently self-justifying equilibrium of high fares, low load factors, and consequently high unit costs. This is not to deprecate the value of

[8] Of course the fact that we have chosen two underutilized airports as the termini in our first two major decisions introducing a policy of multiple permissive entry is itself a reflection of the reality—legally inescapable, we think—that merely proceeding case by case itself involves managing and modulating the process of liberalization.

[9] The "logic of events" includes a critical intervention by the President of the United States. We did attempt in September 1977 to place some limits on the deeply discounted fares offered across the Atlantic by the established, certificated carriers in response to the accentuated competition of charters and Freddie Laker. But the President overturned that attempt of ours and in so doing set us even more rapidly than we otherwise would have been along the path of liberalizing charter rules, and intensified exploration of the possibilities of admitting supplemental carriers—whose very lifeblood is charters—into scheduled service, in order to ensure these price-conscious independents a fair continued opportunity to compete. It also involved the initiation of an aggressive policy of trying to induce foreign governments to let down their barriers to entry by offering them improved access to the *U.S.* market. I have discussed elsewhere this exciting effort to persuade the world's aviation authorities that international trade is not a zero sum game. See my testimony on *U.S.* international aviation negotiations and my 1978a paper.

service competition. The difficulty is that if passengers are presented with no alternative, higher load factor/lower fare offerings, there is not an effective market determination of whether the service offered is too good.

The complete regulator reacts to this dilemma by extending the regulatory net wider, in order to limit these kinds of competition as well: limiting advertising, controlling scheduling and travel agents' commissions, specifying the size of the sandwiches and seats and the charge for inflight movies. The regulatory rule is: each time the dike springs a leak, plug it with one of your fingers; just as dynamic industry will perpetually find ways of opening new holes in the dike, so an ingenious regulator will never run out of fingers.

The efficient way to reverse the process of cost-inflating nonprice rivalry is of course to structure markets competitively, and permit suppliers to vie for customers by reducing their prices. The consequence will be to raise break-even load factors, and, our experience demonstrates, realized load factors as well.

The upshot of these considerations, like the others, was therefore a decision on our part to press forward on both fronts as rapidly as possible—relaxing our previously rigid controls on competition in basic fares,[10] while trying to open up entry rapidly enough to give new, price-competing carriers a fair chance to survive, and to make it irrational for incumbents to try to forestall them by anticipatory, predatory price cuts. The beneficial consequences are already there for anyone to see.

G. *Problem 7: Discriminatory Price Competition*

There are three additional observations that I would like to make about the epidemic of special fares—many of them highly discriminatory—that has broken out during our accelerating process of deregulation.

The first is that many of them are not discriminatory at all, but represent a logical reflection of the varying costs of the various kinds of service this industry provides or is in a position to provide. The marginal opportunity costs, both short and long run, of providing regular coach service—which carries a reasonable probability of a passenger being able to get a seat on relatively short notice on a conveniently scheduled flight, and with no penalty if he fails to show up at flight time—are much higher than of standby service, or of carrying a passenger who volunteers to be bumped from an overbooked flight for sufficient compensation (and we will see more of these, under a new Board order requiring the carriers to seek volunteers before resorting to involuntary bumping); or of charter service—where the passenger accepts the risk of a heavy penalty if he has to cancel out, and of the flight not going out at all, if not enough seats are sold; or of Super-Saver, Budget, or Super-Apexes, the number of which made available on each flight is restricted to the number of seats the carrier estimates would otherwise go out empty, and which are in principle therefore in effect anticipatory standby fares.[11]

In contrast with ordinary standbys, however, these last fares on scheduled service also embody very substantial elements of discrimination. Many of the restrictions on their availability, such as minimum stay requirements, are clearly aimed at confining them to demand elastic customers, and have nothing to do with cost. Moreover, particularly when they were first initiated, they were extremely discriminatory geographically, being available only on particularly competitive, heavily travelled routes. My second observation, however, is that this accentuating price discrimination is symptomatic of the fact that we are still in the transition from tight regulatory cartelization to effective competition: entry is still not free, and until recently the offer of restricted discount fares was the only kind of price competition the Board was willing to permit.

[10]It would be difficult to exaggerate the importance of this change in Board policy, but it would take a separate paper to describe it and analyze its consequences adequately. The landmark decision so far is Domestic Passenger-Fare Level/Fare Structure Policies, PS-80, 43 Fed Reg. 39522 (Sept. 5, 1978).

[11]I have spelled this argument out much more fully in my 1978b paper.

And this leads to the third point, which is that as the process of deregulation proceeds, much of the discrimination will tend to disappear—there are already signs of this happening. Super-Savers, originally available only between New York, Los Angeles and San Francisco, are now available between all major cities in the United States; and you can fly on Super-Apex from many major cities in this country to many major points in Europe—no longer just between New York and London. Texas International's Peanut fares, Continental's Chickenfeed, TWA's No Strings and American's Short Stop are available to all comers in the markets in which they are offered, regardless of size, shape, length of stay, or previous condition of servitude: the only control is that—just like interruptible, off-peak sales of gas and electricity—the number of discounted seats varies from flight to flight, depending upon their timing relative to the system peak; British Caledonian has divided its planes on transatlantic flights into three compartments, with fares in each based upon its own implicit load factor, and therefore on the degree of comfort and ease of obtaining advance reservations that it affords, and with further differentiations based upon the presence or absence of cancellation penalties, stop-over privileges, and circuitous routings—all of them genuine cost-determining variables.

Finally, and most satisfying of all, intensifying competition and the removal of Board prohibitions are at last producing reductions in the basic fares themselves, on a totally nondiscriminatory basis. This process is only just beginning.

III. Epilogue: Who Bears the Burden of Proof?

One of the most fascinating aspects of the public policy disputations I've been participating in during the last four years is the widespread acceptance of the notion that the burden of proof rests always with the advocates of change. That is, even if one is dealing with manifestly irrational, if not idiotic arrangements, the advocate of moving in the direction of rationality is called upon to predict exactly how the process will work out and to prove beyond all doubt that it will work perfectly. In electricity regulation, people who think they will be injured by marginal cost pricing think they fulfill their intellectual responsibilities by a ritualistic incantation of the two magic words, "second" and "best," although some condescend further to enrich the debate by finding some economists willing to contribute some scornful allusions to neoclassical economics. Similarly in air transport, people who profess to be in favor of freer competition nevertheless demand from the advocates of deregulation guarantees that no town will lose service, even temporarily; that no carrier will be subjected to unequal competitive pressures because it may have inherited a less favorable route structure than its rivals; that there will be no wastage of fuel; no excessive entry into any market; no injurious discrimination; no bankruptcies; no loss of seniority rights anywhere; no danger of increased concentration; no impairment of scheduled service. Or they will oppose free entry unless and until the advocates can predict in complete detail how the new pattern of operations will look, while professing to be content to leave the fashioning of the future air system, in its every detail, to the very same *CAB* that stoutly asserts its inability to make those predictions.

The opponents and the faint-hearted importune us to make all our route awards mandatory, exclusive, and rigidly prescribed. The cartelists and protectionists would have us comprehensively prescribe prices, schedules, the size of sandwiches, the pitch of seats, the charge for inflight movies, and travel agents' commissions.

What has been genuinely illuminating to me, in contrast, is how rich a comprehension I have acquired of the distortions of the transition, and how thoroughly I have as a result been converted to the conclusion that the only way to move is fast. The way to minimize the distortions of the transition, I am now thoroughly convinced, is to make the transition as short as possible.

The ultimate consequence is already clearly in sight. The view is growing more and more widespread among the carriers themselves: if the *CAB* no longer provides us with

any protection at all, or exposes us to the distortions of gradual and partial deregulation, wouldn't we be better off with no *CAB* at all? I wish I could say that I had the foresight to have planned it exactly that way!

REFERENCES

H. Averch and L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1052–69.

A. Kahn, Testimony before the Subcommittee on Aviation, House Public Works and Transportation Committee, Sept. 29, 1977.

——, (1978a) "The Changing Environment of International Air Commerce," *Air Law*, Sept. 1978, *3*, 163–74.

——, (1978b) "Deregulation of Air Transport—Getting from Here to There," in *Regulatory Business: The Search for an Optimum*, Inst. Contemporary Studies, San Francisco 1978.

——, "A Paean to Legal Creativity," *Admin. Law Rev.*, forthcoming.

H. C. Simons, "A Positive Program for Laissez Faire," Public Policy Pamphlet No. 15, Chicago 1934; reprinted in his *Economic Policy for a Free Society*, Chicago 1948.

Civil Aeronautics Board, "Improved Authority to Wichita Case," Docket 28848, April 17, 1978.

——, "Chicago Midway Low Fare Route Proceedings," Docket 30277, Order 78-7-40, July 12, 1978.

——, "U.S.-Benelux Exemptions," Docket 32523, Order 78-9-2, Sept. 1, 1978.

——, "Application of World Airways (Guam Exemptions)," Docket 32635, Order 78-9-33, Sept. 7, 1978.

——, "Applications of Colonial Airlines, Inc.," Dockets 30587 and 30591, Order 78-6-183, June 27, 1978; Order 78-8-116, Aug. 22, 1978.

——, "Oakland Service Investigation," Docket 30699, Order 78-4-121, May 30, 1978.

——, "Chicago Midway Expanded Service Proceeding," Docket 33019, Order 78-7-41, July 12, 1978.

Research on Economic Education: Is It Asking the Right Questions?

By BURTON A. WEISBROD*

The division of responsibility between the two papers at this session is a fascinating one. One author has agreed to examine the questions that are being asked in the economics education literature, and the other to examine the answers!

As is so often the case, however, the underlying assumption of separability does not hold. A research question is not a "good" or "bad" question independent of the quality of the answers it is likely to generate. An "exciting" question that is unlikely to yield an answer of substantial value is *not* a good question. Research is a production process in which something called "useful knowledge" is the output. The inputs to this process include both the specification of questions that are important—in the sense that the answers would have great expected value—and the marshaling of resources (i.e., the incurring of costs) to answer the questions.

If the costs of answering all research questions were equal, or were random with respect to the significance of the question, then the separability of the decisions on question *specification* and on question *answering* would be justified. What we probably confront, however, is a less fortuitous set of conditions in which the questions that are most valuable to answer are also the most costly (i.e., difficult). The issue I have been asked to deal with—whether research on economic education is asking the right questions—thus involves implicitly the performance of a benefit-cost analysis on project selection in the area of economic education research. An evaluation is needed of (a) the expected benefits (more precisely, the probability distribution of benefits conditional on answers of various quality), and (b) the expected costs of obtaining answers of each quality.

It is possible, of course, that a particular research question may be a good one in the efficiency sense that the expected costs of researching it are less than what the expected benefits would be *if* the resources devoted to the research were used as productively as possible; yet if the resources were not used so productively, it might fail the allocative efficiency test. Thus, a question could be *potentially* efficient to research but *actually* inefficient. In any event, the "best" questions to research are those for which the excess of the value of the expected answers (benefits) over the expected costs of the research are maximized. Deciding which are the "right" questions to research implies a benefit-cost (efficiency) analysis for the prospective project that is essentially the same as for any other resource-using project, such as in water resources or manpower training. Thus, upon careful scrutiny the imaginative effort by the organizers of this session to break a monstrous evaluation task into two distinct evaluations fails to pass the test of separability.

Despite my conclusion regarding the simultaneity of judgments on which are the right (best) questions to ask and on the costs and quality of expected answers, I shall proceed. In the remainder of this paper, I try to identify the nature of the research questions

*University of Wisconsin-Madison. My thanks go to W. Lee Hansen, Michael Olneck, and Mark Schlesinger for their comments on an earlier draft. This research was supported in part by funds granted to the Institute for Research on Poverty at the University of Wisconsin-Madison by the Department of Health, Education, and Welfare pursuant to the provisions of the Economic Opportunity Act of 1964.

that have been posed in the economic education literature, and the nature of the questions that have *not* been posed. I will comment on whether the overall research program—the set of questions being asked—is what it "should" be, and attempt to point to researchable themes that are likely to have relatively high returns for research in this field.

I. Economic Education and General Education

One basic question is, why study *economic* education at all? What reasons are there to believe that the subject matter of economics is sufficiently special so that the voluminous general literature on teaching and education is not applicable to economics?[1] I have not seen this question posed in the more than 150 papers I have surveyed in the *Journal of Economic Education,* (*JEE*) and in the annual American Economic Association (*AEA*) sessions on economic education.[2] (There are, of course, papers on economics education published elsewhere, but my survey does not extend much beyond these two "official" *AEA* sources.) My point is simple. Is there not a substantial probability—indeed, perhaps not a presumption—that researchers studying economics education are "rediscovering the wheel," posing and answering questions that have been answered previously in the more general research on education? For example, is it not likely that the effect on "learning" of, say, class size, or of the use of teaching assistants rather than more experienced professors, or of individually self-paced approaches rather than a traditional uniform, instructor-paced approach, is similar for all subjects? I do *not* assert that the answer is obvious and affirmative. I only question whether it is a "high priority" research matter to devote substantial resources to general questions of teaching techniques; questions that are not specific to the teaching and learning of economics and that have been studied extensively in other subject matter contexts. It may well be true that, as one economist at an *AEA* session on economic education recently put it, "Educational production functions are at least as interesting as those for hybrid corn" (see Elisabeth Allison, p. 228). Nonetheless, it would not follow that production functions for *economic* education are efficient topics for economists' research.

II. The Production Function for Economic Education

What research has been undertaken in economics education? Most of it is devoted to exploring some portion of the production function for economics education. Of the 159 papers surveyed, I count 102—essentially two-thirds—dealing either with how to define and measure outputs (23 papers), or with the effect on output of various alternative inputs (79 papers). This production function orientation is consistent, however, with the *JEE's* goal as stated inside the front cover: "To promote the teaching and learning of economics in colleges, junior colleges and high schools by sharing knowledge of economic education."

Table 1 presents the 79 input-output oriented papers according to the principal type of input the productivity of which was being studied, and according to the level of schooling. Since each paper was counted only once, while some papers touched on more than one input or school level, the table is an incomplete portrayal of the research foci.

I have classified the independent variables in the production function as capital, labor, students themselves, course content, and instructional methods (ways of combining inputs). An impressive variety of variables have been researched. I cannot judge whether some inputs that have received little or no attention are "worth" studying—for example, the output effects of the time of day that the class is held (but see Rolf Mirus), the color of

[1]This is not to say that there is nothing special about the teaching of economics. Economists typically believe, for example, that people have more misinformation and biases concerning economics than about other subject matters. (Mark Schlesinger pointed this out to me.) Even if this is true (see Kenneth Boulding) the question would remain whether resources devoted to teaching economics should be deployed differently than in other subjects.

[2]For an excellent survey of research on educational production functions, see Eric Hanushek.

TABLE 1—NUMBER OF ARTICLES ON VARIOUS PRODUCTION FUNCTION RELATIONSHIPS FOR ECONOMICS EDUCATION, BY TYPE OF INPUT AND LEVEL OF SCHOOL

Type of Input	Elementary And High School	Junior College	College	Graduate School	Nonschool	All Levels
Capital						
Textbooks	2		2			4
Computers			9			9
Television, slides, etc.		1	5		1	7
						20 (25%)
Labor						
Instructors	6	1	4			11
Graduate Assistants			5			5
Consultants	1					1
						17 (22%)
Students (ability, motivation, family background, other students)	1		4			5 (6%)
Course Content (subject matter)	2	2	7			11 (14%)
Instructional Methods (ways of combining inputs)						
Games and Simulations		1	7			8
Learning contracts, self-paced instruction and programmed learning			11			11
Lectures			1			1
Course evaluations			4			4
Length of course			1			1
Class size			1			1
						26 (33%)
Total	12 (15%)	5 (6%)	61 (77%)	0	1 (1%)	79 (100%)

the classroom walls, or the seating arrangements.

A. *Interaction Effects*

What is probably a more serious omission is the lack of examination of interaction effects among input variables. It seems likely, for example, that a particular type of textbook (input *IA*) when used by graduate teaching assistants (*IIB*) will be more effective for low-ability students (*III*) than they would be for high-ability students. Similarly, games and simulations (input *VA*) may be differentially effective depending on whether instructors (*IIA*) or graduate assistants (*IIB*) are used and depending on the student's intitial level of motivation (*III*).

B. *Limited Scope*

Another striking aspect of Table 1 is the overwhelming emphasis on teaching at the college level (77 percent of the papers). The *JEE* goal, stated above, refers to "colleges, junior colleges and high schools." The scant attention of economics education researchers to high schools and junior colleges is noteworthy, given that half of young people do not go beyond high school, and that those who do go further are increasingly likely to go to a junior college. (Examples of research on economic education in junior colleges are Dennis Weidenaar and Joe Dodson, and Darrell Lewis, Donald Wentworth, and Charles Orvis. For a precollege focus, see Rendigs Fels (1977) and Thomas Duff.) It may or may not be true that the production function findings for the *college* population apply also to the junior colleges and high schools; the issue merits attention. Students' ability and motivation levels (as well as the variances in those levels) vary across the schooling levels; thus, the interactions of these students char-

acteristics with other, conventional inputs will produce, I hypothesize, different output effects depending on the level of school.

The narrow scope of teaching settings on which research has been published is also evident from the dearth of attention to the production function for teaching economics either in graduate schools (see, however, W. Lee Hansen and Robert Decker for models predicting success in graduate economic studies) or in nonschool settings such as in the home via television (see John Coleman) or via popular journalism (magazines and newspapers). How "effective," for example, are the syndicated newspaper columns of writers such as Sylvia Porter, the *Newsweek* columns by Milton Friedman and Paul Samuelson, the articles in magazines such as *Challenge* or *Public Interest*, or in daily newspapers? How effective—and for whom—are the efforts of private firms to provide "economic education" via newspaper advertisements (for example, Mobil Oil on energy issues)? These are unanswered—indeed, unasked—questions. Yet, the vast majority of people have not taken and never will take a formal economics course in any school, and they will be exposed to economics only through such informal media. Thus, the production function for learning economics outside traditional schools seems to warrant substantial exploration—assuming, of course, that economics is worth the opportunity cost of learning it. The omission of nonschool teaching and learning of economics from the *JEE* statement of policy is unfortunate.

C. *Distributional Effects*

I turn next to a related aspect of the production function work: the *distributional* effects of alternative course contents, input combinations, and instructional materials. These have been studied to some extent (for example, Richard Attiyeh and Keith Lumsden; Hansen, Allen Kelley and the author; Fred Thompson); yet, based on the evidence from the general literature on education that a given approach is likely to have substantially different effects on different "types" of students, this dimension seems to deserve more scrutiny. Whatever the mean differential may be between the output effects of different inputs, examination of the variance about the mean may disclose systematic differences among students according to characteristics that are discernible at the outset of a course.

III. Outputs

A. *Goals of Economic Education*

The body of research presented in Table 1 focuses on the productivity of various *inputs;* the dependent variable—output—is generally taken as given, typically in the form of some test score. There is, however substantial other literature—not in the input-output framework—discussing the normative question of how *output* ought to be defined and measured. There are papers that discuss the "usefulness' of a specific output measure, particularly the Test of Understanding in College Economics (*TUCE*) (for example, see Darrell Lewis and Tor Dahl; Fels, 1977). Other concepts and measures of outputs on which papers have been published include changes in student political attitudes (see James Scott and Mitchell Rothman) the students' own judgment of effectiveness (see Kelley); learning "radical" economics (see Richard Edwards and Arthur MacEwen; John Gurley); and developing problem-solving abilities (see Fels, 1973). In addition, the durability or permanence of the effects, as distinguished from measures of effectiveness obtained at completion of the course, has received some attention (see Phillip Saunders; Saunders and G. L. Bach).

Overall, however, the question of what economics education ought to be aiming at—that is, which outputs should be produced—is a question that has received little rigorous analysis. The question of what kind or kinds of "economic education" to produce is a difficult one. Should it be idealogically oriented? Should it provide whatever "buyers" want? Who are the buyers—parents? taxpayers? students? Our custom-

ary consumer sovereignty model appears to be of limited guidance here, given widespread consumer ignorance of the importance of economic knowledge, and given the external benefits from having a population that is more sophisticated in its understanding of economic processes. In economics education, as in many other "professional" markets, buyers are poorly informed regarding product quality. Even if buyers know their objectives, they may know little about the effectiveness of particular activities in achieving those objectives.

My references to "consumer ignorance" and to "external benefits," however, are scarcely more than assertions. I have seen little research that rigorously confronts the question of whether there is a market failure in the economic education market, with too few people studying too little economics or studying the "wrong" economics. The published research either asserts that more economics is good—and presumably is better than some unspecified alternative uses of student time and other resources—or else the research asks the narrower production function question of how effective one type of input is compared to another, without asking whether the output is worth producing. In volume 1 of the *JEE* George Stigler (p. 78) did pose the question, "Why should people be economically literate, rather than musically literate, or historically literate, or chemically literate?" I will resist the temptation to discuss his answer—except to note that musicians, historians, and chemists may see things differently.

B. *Effectiveness vs. Allocative Efficiency*

The domination of a production function emphasis in economic education research has obscured the related issue of the allocative efficiency of alternative input combinations. Many papers have examined the *effectiveness* (productivity) of various inputs, but rarely have the relative *costs* of the inputs been juxtaposed to the relative effectiveness, nor have the measures of effectiveness been translated into *values* of benefits. These questions have seemingly been overlooked or, at least, slighted.

I find it surprising that among the (admittedly small number of) papers confronting the question of how to define the output or outputs of economic education, there has been so little attention to labor market effects in general, and earnings effects in particular. The contrast between the economic education literature and economics of education literature is dramatic. The latter has concentrated, typically within a human capital theoretic framework, on the relationship between education (meaning schooling) and earnings, virtually disregarding the *process* through which educational inputs produce the outputs that have value in the labor market. Another way of saying this is that the economics of education literature has viewed *earnings* as the value of outputs. Meanwhile, the economic education literature has concentrated heavily on the process of converting inputs into outputs in nonpecuniary forms, virtually disregarding the valuation of outputs.

One might have predicted a priori that the economic education literature would have included numerous efforts to assess the labor market value of economics training either directly or indirectly through its effect on, say, the probability of admission to law school. Why the economic education literature and the economics *of* education literature have been so divergent, and whether either, or both, or neither has pursued an "optimal" path are questions which I raise here, but will not pursue far.

C. *Lifetime Effects*

The human capital framework, within which much of the economics of education literature has been cast, has focused research attention on the investment aspect of schooling. The investment emphasis implies a *lifetime* perspective on the outputs of schooling. By sharp contrast, the economic education literature has concentrated overwhelmingly on the *immediate* outputs, those measured at the completion of the course. As pointed out

above, there have been a few noteworthy exceptions in which the durability of outputs has been considered, though even these have involved a horizon of only a few years or so (see Saunders; Saunders and Bach). It may well be exceedingly difficult to measure lifetime effects of exposure to economics, and this may explain the lack of attention to this question in the literature. (This would illustrate the interrelatedness of the "do-ability" of research and the formulation of research questions.) But the fact remains that little effort has been devoted to the measurement of lifetime effects.

IV. Incentive Structures

Another underresearched area is the nature of incentive structures facing teachers and admininstrators. Assume that 1) the production function research disclosed that certain inputs are more effective than others, 2) consensus was reached on appropriate measures of outputs (i.e., effectiveness), and 3) outputs and inputs were valued and showed positive net benefits from a change in current teaching practices. Would the changes occur? Are there incentives sufficient to encourage changes that are efficient (granting that such changes can be identified with reasonable confidence)?

These questions, it might be argued, transcend economic education. It would seem, however, that the responsiveness of teachers and administrators of economic education programs may or may not be the same as for those in noneconomics areas; at least this hypothesis cannot be ruled out, any more than can the hypothesis that variation in class size, or in the effectiveness of teaching assistants, or the use of television instruction differs as between economics and other subject areas.

The nature of incentives confronting teachers—of economics or of anything else, and at various levels of schooling—has received scant attention. There are possible incentives for instructors (a) to learn which changes are efficient, and (b) to make those changes. (On the latter point, studies of salary determination at universities, see, for example, John Siegfried and Kenneth White. James Koch and John Chizmar have shed some light on the financial returns to scholarly research, teaching, and other uses of faculty time.) It is arguable that little is to be gained from research on *how* to "improve" teaching if the *incentives* to adopt improved methods are weak. It is also arguable, on the other hand, that incentives are weaker than they might be because there is so little agreement as to what constitutes efficient teaching; this, after all, involves the specification of goals in operational terms and the adoption of value weights for the multiple goals that surely exist. Thus, understanding *goals and weights* is one part of the research agenda for efficient innovation in education.

In any analysis of incentives in education the relationship between private costs and social costs (or returns) is likely to be crucial. As an illustration, consider the case of an economics instructor who is free (although many are not) to select any undergraduate textbook, and that a new textbook appears on the market. There may well be little incentive (financial, professional, or any other) to read the new textbook carefully enough to determine whether it is superior to the one being used; this, however, is not my principal point. What if the instructor knew—costlessly and with certainty—that the new book was "more effective" for all of his students; what would be the private and social costs and benefits of adopting the new book? Of course, more effective need not imply "more efficient."

From the students' viewpoint, the new book would presumably be preferred if it were more effective. Such a preference, in turn, embodies two deeper assumptions: (a) the similarity of student goals and of faculty goals for students, and (b) the absence of higher costs (time, effort, money) for using the new book that offset the benefits of increased learning.

Note, however, that while the student must incur the cost of reading whatever textbook is chosen—an essentially fixed cost—the faculty person bears an increased real social cost of changing, since he or she has lecture notes keyed to a textbook that has already been

read. With the benefits of change accruing to students while the costs are borne by faculty, the likelihood of market failure is substantial.

The market failure would disappear, however, if the instructor internalized the students' benefits. This might appear to be the case if the instructor acted as an idealized "professional"—that is, acted as the consumer's agent for maximizing the consumer's (student's) utility. Education is an example of a commodity—like medical care and legal representation—in which consumers are aware of their inability to judge quality, and so they place trust in the professional to act in their best interest. Even if the instructor were to behave, however, so as to maximize not his or her own utility but that of students (or parents, or taxpayers), it would not follow that efficient resource allocation would result. The reason is that the cost of switching textbooks (or, in general, of changing anything in the teaching process) is a real cost; if it were to be disregarded—as would be the case if the instructor were to act so as to maximize the consumer's utility—the result would be excessive change.

The market failure would also disappear if the reward structure were such that the instructor's pay were an appropriate function of the "value-added." Then, if students learned more from the new text, the instructor—acting in self-interest—would weigh the costs of changing books against the benefit, and would choose accordingly. Ideally, the rewards would be commensurate with the student benefits, and so—assuming away real external effects and other market imperfections—the instructor would be confronted with the real costs and benefits of change. The problems of developing such a reward system are doubtless great. It does not follow, however, that they are not researchable.

These remarks have been abstract; meat must be put on the analytic bones. I hope that the next time the economic education literature is surveyed there will be found more papers exploring incentives for innovation and efficiency—in both positive and normative dimensions.

V. Concluding Remarks

It is all too simple to find questions that one would like *other* researchers to tackle, as I have done here. Thus, I should close by reiterating my claim made at the outset that the selection of optimal research questions is, in principle, a matter of weighing benefits and costs, of comparing the value of having answers to the costs of obtaining them. If the costs are sufficiently high, it would be inefficient to research questions that seem important. Some of the questions to which I have pointed probably fail such a benefit-cost efficiency test, and so have received, quite wisely, little research attention; other questions, however, may pass it—at least for some researchers—and so merit more study. Once more we can conclude that "more research is needed."

REFERENCES

E. Allison, "Three Years of Self-Paced Teaching in Introductory Economics at Harvard," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 222–28.

R. Attiyeh and K. Lumsden, "Modern Myths in Teaching Economics: U.K. Experience," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 429–433.

K. Boulding, "Some Observations on the Learning of Economics," *Amer. Econ. Rev Proc.*, May 1975, *65*, 428–30.

J. Coleman, "Economic Literacy: What Role for Television?," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 645–52.

R. L. Decker, "Success and Attrition Characteristics in Graduate Studies," *J. Econ. Educ.*, Spring 1973, *4*, 130–37.

T. B. Duff, "Basic Economic Concepts in the High School Curriculum," *J. Econ. Educ.*, Fall 1971, *3*, 5–10.

R. C. Edwards and A. MacEwen, "A Radical Approach to Economics," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 352–63.

R. Fels, " A New Test of Understanding in College Economics," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 660–66.

———, "Developing Independent Problem-Solving Ability in Elementary Economics," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 403–07.

———, "What Economics is Most Important to Teach: The Hansen Committee Report," *Amer. Econ. Rev. Proc.*, Feb. 1977, *67*, 101–04.

J. **Gurley**, "The Principles Course: What Should be in it and Where Should it be Going?," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 431–33.

W. L. Hansen, "Prediction of Graduate Performance in Economics," *J. Econ. Educ.*, Fall 1971, *3*, 49–53.

———, **A. C. Kelley and B. A. Weisbrod**, "Economic Efficiency and the Distribution of Benefits from College Instructions," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 364–69.

E. Hanushek, "A Reader's Guide to Educational Production Functions," work. paper no. 798, Instit. Soc. Pol. Stud., Yale Univ. 1977.

A. C. Kelley, "Uses and Abuses of Course Evaluations as Measures of Educational Output," *J. Econ. Educ.*, Fall 1972, *4*, 13–18.

J. Koch and J. Chizmar, "The Influence of Teaching and Other Factors upon Absolute Salaries and Salary Increments at Illinois State University," *J. Econ. Educ.*, Fall 1973, *5*, 27–34.

D. R. Lewis and T. Dahl, "The Test of Understanding in College Economics and its Construct Validity," *J. Econ. Educ.*, Spring 1971, *2*, 155–66.

———, **D. R. Wentworth, and C. C. Orvis**, "Economics in the Junior Colleges: Terminal or Transfer?," *J. Econ. Educ.*, Spring 1973, *4*, 100–10.

R. Mirus, "Some Implications of Student Evaluation of Teachers," *J. Econ. Educ.*, Fall 1973, *5*, 35–37.

P. Saunders, "Does High School Economics have a Lasting Impact?," *J. Econ. Educ.*, Fall 1970, *2*, 39–55.

——— **and G. L. Bach**, "The Lasting Effects of an Introductory Course: An Exploratory Study," *J. Econ. Educ.*, Spring 1970, *1*, 143–49.

J. H. Scott Jr., and M. P. Rothman, "The Effect of an Introductory Economics Course on Student Political Attitudes," *J. Econ. Educ.*, Spring 1975, *6*, 107–12.

J. Siegfried, and K. White, "Financial Rewards to Teaching and Research," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 309–15.

G. Stigler, "The Case, if any, for Economic Education," *J. Econ. Educ.*, Spring 1970, *1*, 77–84.

F. A. Thompson, "Problems and Prospects of Economics Education in Community Junior Colleges," *J. Econ. Educ.*, Fall 1970, *2*, 31–38.

D. J. Weidenaar and J. A. Dodson Jr., "The Effectiveness of Economics Instruction in Two-Year Colleges," *J. Econ. Educ.*, Fall 1972, *4*, 5–12.

Research on Economic Education: How Well is It Answering the Questions Asked?

By THOMAS JOHNSON*

In reviewing the research on economic education I have sensed a growing maturity in the use of statistical and econometric techniques. However, unlike the closely related research on the economics of education, this research has not been prolific in providing examples and problems for econometric research. In this review I will discuss four econometric issues which have surfaced in the last ten years of research on economic education. Perhaps this will provide some guidance to future researchers in economic education and also reach an audience of econometricians who will find some intriguing examples and problems in this field. The issues which I will discuss are: 1) structural equations and errors in variables, 2) limited and qualitative dependent variables, 3) multicollinearity, and 4) other concerns.

I. Structural Equations

The paper by David Ramsett, Jerry Johnson and Curtis Adams has a definite simultaneous equations structure problem in the techniques employed. The two dependent variables analyzed are Post-*TUCE* (Test of Understanding in College Economics) and Post-Attitude, with each appearing in the equation for the other. Despite this obvious simultaneity, ordinary least squares (*OLS*) is used with no discussion of simultaneous equation bias. Another example of an actual simultaneous equation structural problem is the paper by John Soper (1976) in which he attempts a two-stage procedure. However, rather than creating instrumental variables as in the standard two-stage least squares, Soper enters the residuals from the first stage into his second stage regressions. Craig Swan has recently critiqued this procedure more extensively than space permits in this paper. In summary, Swan notes that if a simultaneous equations bias exists in Soper's original equations (which are recursive), then the same problems exist in the second stage of Soper's procedure. A third example of an actual simultaneous equation structure is the paper by William Becker and Michael Salemi which I will discuss more fully below. The last example is the paper by Daniel Thornton and George Vredeveld in which they estimate a recursive system using ordinary least squares. While the use of *OLS* may be optimal, no analysis is provided to support this choice.

My conclusion is that in those (rare) published papers where a simultaneous equation structure has been formulated, the original researchers have used statistical techniques which do not answer well the questions raised by their choice of structure. An exception is the recent paper by Elisabeth Allison which does employ adequate methods to estimate a simultaneous structure with student effort, achievement and satisfaction as endogenous variables. It is missed opportunities rather than explicit errors which give major importance to this issue. For example, opportunities for use of structural equation modeling are found in the papers by Allen Kelley (1972, 1975), Robert Highsmith (1974), Andrew Nappi, Lewis Karstensson and Richard Vedder, Howard Tuckman, and Becker and Salemi. Richard McKenzie has already noted the possibilities for structural analysis by Kelley (1972). Kelley analyzes measures of course evaluations and professor evaluations by students using separate equations and *OLS*, although one would think that the attitudes would be correlated with the evaluations.

Nappi includes a Flanagan Test of General Ability (given before the course) in explaining the postcourse *TEU* score. The methods used

*North Carolina State University. I wish to thank Ronald A. Schrimper, David T. Barker, and Zvi Griliches for helpful comments while absolving them of any liability for the final product.

are adequate for estimation of direct effects *if* the Flanagan test was an errorless measure of ability. But might there not be indirect effects of ability? With how much error does the Flanagan test measure the relevant variables? These questions are very similar to those asked by Zvi Griliches and William Mason. Recent developments in multivariate statistical methods (see Karl Jöreskog; Jöreskog and Arthur Goldberger; Gary Chamberlain, 1977a, b; Chamberlain and Griliches) seem to be especially designed to answer questions of this kind. Besides combating simultaneous equations bias, use of these techniques may increase the precision of parameter estimates when variables are unobservable or measured with error.

Karstensson and Vedder, and Kelley (1975) miss particularly good opportunities to employ structural analysis. Karstensson and Vedder analyze the determinants of grades in a beginning economics course and the change in attitude toward economics in separate equations using ordinary least squares. Surely these two variables are likely to affect each other. Kelley analyzes course performance and attendance separately. Kelly does include dummy variables to designate sophomores and upperclassmen which might be the most logical way of including existing human capital other than specific skills measured by scores on the Scholastic Aptitude Test or other courses. Since these are not precise measures of existing human capital, the statistical techniques of Jöreskog might help to overcome the problem of errors in variables.

Tuckman analyzes, in separate equations, five measures of teacher effectiveness. These are 1) a ten question version of *TUCE*, 2) an attitudes test, 3) course grade; 4) an index of student interest; 5) willingness to take more economics courses. He begins to move toward structural analysis by investigating zero-order correlations between these variables. He also includes class as a single variable in interval form (freshman = 1 . . . senior = 4). Formulating an explicit model and further empirical techniques derived from Jöreskog's work might be useful in devising an index of teacher effectivenss.

Becker and Salemi do explicity recognize simultaneous equation bias and attempt to deal with the problem of ability or aptitude (*APT*) being a unobservable exogenous variable. Their two-equation structural system is:

$$DTUCE = a + b\,APT + c\,\text{Situation} + d\,\text{Time} + u$$

$$\text{pre-}TUCE = \gamma + \theta\,APT + \nu$$

They then eliminate the unobservable *APT* obtaining the "reduced form"

$$DTUCE = a' + b'\,\text{pre-}TUCE + c'\,\text{Situation} + d\,\text{Time} + w$$

which is then estimated using two-stage least squares to obtain an instrument for pre-*TUCE*. However, the properties of this procedure are neither referenced nor investigated. The preferred procedure would be to seek an identified structure and then to proceed with the analysis of unobservable variables.

To summarize, the field of research on economic education is developing to the point where explicit use of simultaneous equations structures is frequently indicated. The use of recently developed statistical techniques for analysis of structures with unobservable variables and errors in variables should be as useful in this research as they have been in other areas.

II. Limited and Qualitative Dependent Variables

Much of the research in economic education seeks explanations for the level of performance on a test of understanding such as the Test of Understanding in College Economics (*TUCE*) which employs thirty-three multiple choice items. Naturally, there are limits to the performance of a student on these "objective" type tests. The lower limit would seem to have little effect on test results since on a question with four alternatives there is a .25 probability of getting the correct answer by completely random choice. However, the possibility of a ceiling would seem to limit the measurement of performance of very able students. The study by Frank Gery has been frequently cited in the literature and

concludes that floor effects may be of more importance than ceiling effects in biasing the estimates of regression parameters. Unfortunately, the methods by which Gery concludes that the "gap-closing" model is less defective are flawed.

Gery compares four alternative models, each characterized by a different specification of the dependent variable. The four dependent variables combine the Post-*TUCE* and Pre-*TUCE* scores in the following ways: 1) *POST*, 2) *POST* − *PRE*, 3) (*POST* − *PRE*)/*PRE*, and 4) the gap-closing model in which the dependent variable is (*POST* − *PRE*)/(33 − *PRE*). Each is regressed on *PRE* as the independent variable. Since the dependent variables are different, the R^2s are not comparable between forms. This is most clearly seen by noting that forms 1) and 2) are essentially the same (only the coefficient on *PRE* is reduced by exactly 1.0), but they yield different R^2s. Also, none of these models includes all of the others as a special case, so that usual parametric statistical techniques do not provide tests of hypotheses for choosing among these forms.

Similar concerns with the effects of limited dependent variables and use of the gap-closing model appear in Soper (1973), Soper and Thornton, Highsmith (1976), and Becker and Salemi. Only Becker and Salemi go further and use a logistic formulation with the gap-closing form of the dependent variable. Considering the central importance of the *TUCE* scores in the analysis of economic education, much more careful work should be done investigating the effects of limits in the variable. As the variable approaches a limit, the possible error is truncated by the presence of the limit and the mean of the error is not zero. If ordinary least squares regression is used, this nonzero mean of the error near the limit can cause biased estimates of parameters. The techniques used by Takeshi Amemiya seem particularly appropriate. James Heckman offers a simplified approach and uses it in a labor market context.

The instances of use of 0–1 or qualitative dependent variables have been rather rare in the literature reviewed. John Lloyd analyzes Yes-No and three-way responses using χ^2 contingency table techniques. Wallace Oates and Richard Quandt use the *Z* test to try to detect differences in proportions of economics majors attracted by faculty and graduate student instructors of principles. More recently, Robert Decker performs *t*-tests of ratios of success and attrition in graduate studies. Also, William Barnes uses a linear discriminant function to explain student acquisition of information and interprets the results as probabilities. Several other studies perform *t*-tests of differences between means of characteristics of groups. However, I have found no examples of use of logit analysis with a dichotomous dependent variable, multiple independent variables, and observations on individuals. The logit procedure developed by Marc Nerlove and S. James Press makes this type of analysis very economical for the few hundred observations and relatively small number of independent variables typical of studies of economic education. It is also now feasible, although much more costly, to analyze situations with several categories of outcome. Daniel McFadden (1976, 1978) provides very useful discussions of these techniques.

III. Multicollinearity

Recently concern for effects of multicollinearity have become more prominent in *The Journal of Economic Education*. The issue of multicollinearity surfaces in papers by Norman Wood and Charles DeLorme, Soper and Thornton, in the debate between Becker, Soper (1976), and Highsmith (1976), and most recently in Swan. The approach in each of these papers is to look for conditions of near singularity of the $X'X$ design matrix. Wood and DeLorme, and Swan do not mention specific statistical tests: the rest rely upon the test presented by Donald Farrar and Robert Glauber. Highsmith (1976) refers to this as "a now standard test." This claim of a "standard" misses the criticism of Yoel Haitovsky who pointed out that it would be better *not* to test for deviations from orthogonality, but rather to test for the difference of $|X'X|$ from zero. Additional criticisms of the Farrar and Glauber test have been presented in notes

by T. Krishna Kumar, C. Robert Wichers, and John O'Hagan and Brendan McCabe. Soper and Thornton (p. 85) make the error of claiming that multicollinearity can cause estimates of coefficients to be biased and inconsistent. However, it is the imposition of restrictions, such as dropping variables, to combat multicollinearity which may create bias.

As noted by G. S. Maddala, and the author and T. D. Wallace, multicollinearity is not a problem of the population from which the X's are drawn. It is a problem of the sample of X's used and the covariation between these observations and the observations on the dependent variable. While Farrar and Glauber's test is inappropriate, their advice to seek more information when a problem of multicollinearity is suspected is sound. If one is finally reduced to mere statistical trickery, an application of ridge regression techniques may be appropriate. A survey of these techniques is available in a recent paper by Hrishikesh Vinod.

IV. Other Concerns

It is noted in Section II that Gery compares equations which were not properly nested for hypothesis testing. Similar problems are encountered when Soper and Thornton search for non-linear forms but conclude that their linear, additive form is the most effective functional form (p. 89). Although Soper and Thornton do not explain their method of search, the context indicates that they follow Gery's method. On the other hand, Steven Cox is very careful to show the equations tried and references an appropriate nonparametric test.

Even with the unique opportunity to collect data from well-designed experiments, researchers in economic education should consider the problems posed by self-selection and censoring in samples. Even if students are randomly assigned to treatment and control groups, their option to drop or transfer may pose statistical problems. David Wise has analyzed similar problems in the analysis of income maintenance experiments. The computational techniques are only moderately costly in computer time, but the algorithms may be relatively difficult to implement if a specialist is not available.

Finally, let me raise a question which has been bothering me and for which I do not have an answer. Are the measures of achievement or attitude which we seek to explain cardinal, or are they only ordinal representatives of the objects of concern? This question seems particularly important in the works by Donald Davidson and John Kilgore, W. Lee Hansen, James Koch and John Chizmar, and Ramsett, Johnson, and Adams, although almost every paper with data pose this question. It is obvious from a little reading in the testing literature that this is not a novel question. However, a survey of how this problem might effect the research in the economics of education might be helpful to future researchers and to those who would apply the results. Or was this one laid to rest before I came in?

REFERENCES

E. Allison, "Educational Production Function for an Introductory Economics Course," disc. paper no. 545, Harvard Inst. Econ. Res., Harvard Univ., Apr. 1977.

T. Amemiya, "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, Nov. 1973, *41*, 997–1016.

W. F. Barnes, "Test Information: An Application of the Economics of Search," *J. Econ. Educ.*, Fall 1975, *7*, 28–33.

W. E. Becker, Jr., "Programmed Instruction in Large-Lecture Courses: A Technical Comment," *J. Econ. Educ.*, Fall 1976, *8*, 38–40.

——— **and M. K. Salemi**, "The Learning and Cost Effectiveness of AVTS Supplemented Instruction: Specification of Learning Models," *J. Econ. Educ.*, Spring 1977, *8*, 77–92.

Gary Chamberlain, (1977a) "An Instrumental Variable Interpretation of Identification in Variance-Components and MIMIC Models," in Paul Taubman, ed., *Kinometrics: Determinants of Socioeconomic Success Within and Between Families*, Amsterdam

1977, 235–54.

———, (1976) "Education, Income, and Ability Revisited," *J. Econometrics*, Mar. 1977, *5*, 241–57.

——— **and Zvi Griliches**, "More on Brothers," in Paul Taubman, ed., *Kinometrics: Determinants of Socioeconomic Success Within and Between Families*, Amsterdam 1977, 97–124.

S. R. Cox, "Computer-Assisted Instruction and Student Performance in Macroeconomic Principles," *J. Econ. Educ.*, Fall 1974, *6*, 29–37.

D. G. Davidson and J. H. Kilgore, "A Model for Evaluating the Effectiveness of Economic Education in Primary Grades," *J. Econ. Educ.*, Fall 1971, *3*, 17–25.

R. L. Decker, "Success and Attrition Characteristics in Graduate Studies," *J. Econ Educ.*, Spring 1973, *4*, 131–37.

D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *Rev. Econ. Statist.*, Feb. 1967, *49*, 92–107.

F. W. Gery, "Is There a Ceiling Effect to the Test of Understanding in College Economics?," in Arthur L. Welsh, ed., *Research Papers in Economic Education*, New York 1972, 35–49.

Z. Griliches and W. M. Mason, "Education, Income, and Ability," in Arthur S. Goldberger and Otis Dudley Duncan, eds., *Structural Equation Models in the Social Sciences*, New York 1973, 285–316.

Y. Haitovsky, "Multicollinearity in Regression Analysis: Comment," *Rev. Econ. Statist.*, Nov. 1969, *51*, 486–89.

W. L. Hansen, "Prediction of Graduate Performance in Economics," *J. Econ. Educ.*, Fall 1971, *3*, 49–53.

J. J. Heckman, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals Econ. Soc. Measure.*, Fall 1976, *5*, 475–92.

R. Highsmith, "A Study to Measure the Impact of In-Service Institutes on the Students of Teachers Who Have Participated," *J. Econ. Educ.*, Spring 1974, *5*, 77–81.

———, "Second Generation Research in Economic Education: Comment," *J. Econ. Educ.*, Fall 1976, *8*, 48–51.

T. Johnson and T. D. Wallace, "Multicollinearity and No-FLIRP," *J. Econ. Develop.*, July 1976, *1*, 29–35.

K. G. Jöreskog, "A General Method for Estimating a Linear Structural Equation System," in Arthur S. Goldberger and Otis Dudley Duncan, eds., *Structural Equation Models in the Social Sciences*, New York 1973, 85–112.

——— **and A. S. Goldberger**, "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *J. Amer. Statist. Assn.*, Sept. 1975, *70*, 631–39.

L. Karstensson and R. K. Vedder, "A Note on Attitude as a Factor in Learning Economics," *J. Econ. Educ.*, Spring 1974, *5*, 109–11.

A. C. Kelley, "Uses and Abuses of Course Evaluations as a Measure of Educational Output," *J. Econ. Educ.*, Fall 1972, *4*, 13–18.

———, "The Student as a Utility Maximizer," *J. Econ. Educ.*, Spring 1975, *6*, 82–92.

J. V. Koch and J. F. Chizmar, "The Influence of Teaching and Other Factors upon Absolute Salaries and Salary Increments at Illinois," *J. Econ. Educ.*, Fall 1973, *5*, 27–34.

T. K. Kumar, "Multicollinearity in Regression Analysis," *Rev. Econ. Statist.*, Aug. 1975, *57*, 365–66.

J. W. Lloyd, "Role Playing, Collective Bargaining, and the Measurement of Attitude Change," *J. Econ. Educ.*, Spring 1970, *1*, 104–10.

D. McFadden, "Quantal Choice Analysis: A Survey," *Annals Econ. Soc. Measure.*, Fall 1976, *5*, 363–90.

———, "On the Use of Probabilistic Choice Models in Economics," paper presented at the meeting of the American Economic Association, Chicago, Aug. 30, 1978.

R. B. McKenzie, "The Economic Effects of Grade Inflation on Instructor Evaluations: A Theoretical Approach," *J. Econ. Educ.*, Spring 1975, *6*, 99–106.

G. S. Maddala, *Econometrics*, New York 1977,

183–94.

A. T. **Nappi,** "An Evaluation of Award-Winning Elementary Teaching Materials from the Kazanjian Program," *J. Econ. Educ.*, Spring 1974, *5*, 82–88.

M. **Nerlove and S. J. Press,** "Multivariate Log-Linear Probability Models for the Analysis of Qualitative Data," disc. paper no. 1, Center Statist. Probability, Northwestern Univ., June 1976.

W. E. **Oates and R. E. Quandt,** "The Effectiveness of Graduate Students as Teachers of Principles of Economics," *J. Econ. Educ.*, Spring 1970, *1*, 130–38.

J. **O'Hagan and B. McCabe,** "Tests for the Severity of Multicollinearity in Regression Analysis: A Comment," *Rev. Econ. Statist.*, Aug. 1975, *57*, 368–70.

D. E. **Ramsett, J. D. Johnson, and C. Adams,** "Some Evidence on the Value of Instructors in Teaching Economic Principles," *J. Econ. Educ.*, Fall 1973, *5*, 57–62.

J. C. **Soper,** "Programmed Instruction in Large-Lecture Courses," *J. Econ. Educ.*, Spring 1973, *4*, 125–29.

——, "Second Generation Research in Economic Education: Problems of Specification and Interdependence," *J. Econ. Educ*, Fall 1976, *8*, 40–48.

—— **and Richard M. Thornton,** "Self-Paced Economics Instruction: A Large-Scale Disaggregated Evaluation," *J. Econ. Educ.*, Spring 1976, *7*, 81–91.

C. **Swan,** "Comments on the Problems of Specification and Interdependence in a Set of Learning Equations," *J. Econ. Educ.*, Spring 1978, *9*, 81–86.

D. L. **Thornton and G. M. Vredeveld,** "In-Service Education and Its Effect on Secondary Students: A New Approach," *J. Econ. Educ.*, Spring 1977, *8*, 93–99.

H. P. **Tuckman,** "Teacher Effectiveness and Student Performance," *J. Econ. Educ.*, Fall 1975, *7*, 34–39.

H. D. **Vinod,** "A Survey of Ridge Regression and Related Techniques for Improvements Over Ordinary Least Squares," *Rev. Econ. Statist.*, Feb. 1978, *60*, 121–31.

C. R. **Wichers,** "The Detection of Multicollinearity: A Comment," *Rev. Econ. Statist.*, Aug. 1975, *57*, 366–68.

D. **Wise,** "Sample Selection, Attrition and Income Maintenance Experiments," paper presented at the meeting of the American Economic Association, Chicago, Aug. 29, 1978.

N. J. **Wood and C. D. DeLorme, Jr.,** "An Investigation of the Relationship among Teaching Evaluation, Research and Ability," *J. Econ. Educ.*, Spring 1976, *7*, 77–80.

A Confusion of Economists?

By J. R. KEARL, CLAYNE L. POPE, GORDON C. WHITING, AND LARRY T. WIMMER*

Following the 1977 meetings of the American Economic Association, *Business Week*, in an editorial entitled "The Furniture Movers," suggested "[t]here was no evidence of either humility or competence at the AEA meeting. Nor did any economist or group of economists offer anything resembling a new idea for addressing the major policy dilemma of the industrial West . . . [i]nstead, the sessions were dominated by papers seeking to refine methodologies that already have been proven ineffective . . . like their counterparts, the moving men, economists collect money and hours for pushing the furniture around . . ." This observation reflects an often heard theme about what the economics profession has to communicate—that it is overly concerned with questions which are obscure, arcane, irrelevant, theoretical, academic, etc., and which are consequently "unimportant."

A second, common, perception of economics and economists is that there is widespread and serious disagreement about "important" issues and hence that economists can contribute little to analysis, solutions, or understanding of these issues. Indeed, this perception is part of our folklore: "If parliament were to ask six economists for an opinion, seven answers would come back—two, no doubt, from the volatile Mr. Keynes" (see Paul Samuelson, p. 1628).

In part, the profession is a source of this perception of disagreement. For example, Edwin Dale noted that Britain's decision to join the common market generated the following exchange of letters to the *London Times*. One letter with 154 signatures of economists arguing that "the economic effects of joining the common market . . . are more likely to be unfavorable than favorable to Britain," was followed by a second letter from 142 economists which concluded "the economic effects of joining the common market . . . are more likely to be favorable than unfavorable to Britain" (Oct. 22, 1971). (A Mr. Peter Sieber then wrote to suggest that "the economic effects of economists . . . are more likely to be unfavorable than favorable to Britain" (Oct. 25, 1971).) Dale commented that the United States had witnessed a battle between monetarists and fiscalists "while the economy was going to pot" and, he continued, "[a]bove all, possibly unfairly, we have had a rise in skepticism about what economists can tell us . . ." He concluded by asking, "are we seeing the decline and fall of the economists' empire?" Similarly, *Business Week* suggested that it was observing the "intellectual bankruptcy of the profession."

These perceptions of irrelevance and/or disagreement may, unfortunately, be used by policymakers to justify the abandonment of analysis and the adoption of simplistic and perhaps superficial answers to complex problems where potential insights might be obtained with economic analysis. The image of irrelevance can only be dispelled by experience with serious applied analysis. However, the image of widespread disagreement may not be a misperception at all. Whether it is or not, and the possible sources of the disagreement are considered in this paper.

I. Hypotheses

In thinking about the problem of widespread perceived disagreement among econo-

*Kearl and Pope are associate professors of economics, Wimmer is a professor of economics, and Whiting is an associate professor of communications and social science, all at Brigham Young University.

mists, we outlined two *ex ante* hypotheses:

1) There would be widespread consensus about questions which were more clearly of a micro-economic nature and less consensus about those with a macro-economic orientation.

2) There would be greater consensus about questions of what *can* be done than about those concerned with what *ought* to be done. This can-ought to split we thought of as loosely analogous to a focus on positive issues (where theory might give direction) vs. normative issues (where value judgments might be important.)

In addition to an interest in the degree of consensus, we wanted to examine common response patterns across issues by groups of individuals. Others have also been concerned about the problems suggested in our introductory comments and have suggested alternative hypotheses about their source. Samuelson, for example, suggests that policy statements usually involve forecasting which is both difficult and an art form where esthetic sensibilities vary. He also argues that ethical ends dominate scientific judgments. There is not, however, much consensus even here: Milton Friedman argues that the lack of agreement is largely a result of differing scientific evaluations, not value judgment per se. He suggests other reasons including product differentiation, a natural desire to talk about differences of opinion and different rates of time preference. (According to Friedmen, advocates of limited government and free markets have a low rate of time preference relative to the advocates of government intervention.)

II. The Survey and Questionnaire

To investigate the questions raised above, a two-page mail-return format questionnaire was sent to a stratified random sample of 600 *U.S.* based economists selected from the 1974 *Directory of Members* of the American Economic Association. The strata included:

1. A random sample of 100 economists from among the full professors of economics in seven leading graduate programs.
2. A random sample of 200 other economists with academic appointments.
3. A random sample of 150 economists employed in government positions.
4. A random sample of 150 economists employed in the private nonacademic sector.

Each recipient was asked to indicate general agreement, agreement with provisions or general disagreement with each of thirty statements. A written response about the effectiveness and sophistication of economic journalism also was requested. While we shall not report upon these comments in detail, the 72 individuals who chose to respond felt (predictably) that the source of confusion was not with professional economists but with journalists. Economic journalism was described as inaccurate, biased, distorted, sensationalized, and unsophisticated. A fair summary of the comments was provided by one respondent who concluded, "All economic journalists must have taken introductory economics on a pass-fail basis."

We were concerned that hostility toward questionnaires among economists might limit the response. We were delighted with a better than 33 percent return; 211 individuals responded: 25 from the first stratum, 81 from the second, 48 from the third, and 57 from the last. (A *chi*-square test of those differences does not reach significance.)

We did not pretest the questionnaire, hence there are some propositions which respondents found inappropriate. While we did not solicit comments on propositions, many respondents chose to comment anyway. (Proposition numbers shown in italics throughout the paper are as defined in Table 1.) Proposition *11* (reduce defense spending) stimulated the heaviest response; only proposition *5* (flexible exchange rates) was without any written comment. The most serious criticisms concerned proposition *26* (Phillips curve)—respondents considered it either meaningless or misleading as written. Affirmative consensus was predicted if it had been worded, "A reduction in unemployment tends to produce a higher rate of inflation."

Table 1 reproduces the propositions and summarizes the responses. Table 2 provides a response breakdown by occupational stratum, assigning a score of 3 to agree, 2 to agree with provisions, and 1 to disagree.

TABLE 1—QUESTIONNAIRE AND RESPONSES
(Shown in Percent)

Propositions	Generally Agree	Agree with Provisions	Generally Disagree
1. Tariffs and import quotas reduce general economic welfare	81	16	3
2. The government should be an employer of last resort and initiate a guaranteed job program	26	27	47
3. The money supply is a more important target than interest rates for monetary policy	48	23	29
4. Cash payments are superior to transfers-in-kind	68	24	8
5. Flexible exchange rates offer an effective international monetary arrangement	61	34	5
6. The "Corporate State," as depicted by Galbraith, accurately describes the context and structure of the *U.S.* economy	18	34	48
7. A minimum wage increases unemployment among young and unskilled workers	68	22	10
8. The government should index the income tax rate structure for inflation	41	27	32
9. Fiscal policy has a significant stimulative impact on a less than fully employed economy	65	27	8
10. The distribution of income in the United States should be more equal	40	31	29
11. National defense expenditures should be reduced from the present level	36	30	34
12. Antitrust laws should be used vigorously to reduce monopoly power from its current level	49	36	15
13. Inflation is primarily a monetary phenomenon	27	30	43
14. The government should restructure the welfare system along lines of a "negative income tax"	58	34	8
15. Wage-price controls should be used to control inflation	6	22	72
16. A ceiling on rents reduces the quantity and quality of housing available	78	20	2
17. The Fed should be instructed to increase the money supply at a fixed rate	14	25	61
18. Effluent taxes represent a better approach to pollution control than imposition of pollution ceilings	50	31	19
19. The government should issue an inflation indexed security	33	25	42
20. The level of government spending should be reduced (disregarding expenditures for stabilization)	34	23	43
21. The Fed has the capacity to achieve a constant rate of growth of the money supply if it so desired	35	41	34
22. Reducing the regulatory power of the *ICC*, *CAB* et al would improve the efficiency of the *U.S.* economy	47	31	22
23. The federal budget should be balanced over the business cycle rather than yearly	53	30	17
24. The fundamental cause of the rise in oil prices of the past three years is the monopoly power of the large oil companies	11	14	75
25. The redistribution of income is a legitimate role for government in the context of the *U.S.* economy	52	29	19
26. In the short run, unemployment can be reduced by increasing the rate of inflation	31	33	36
27. The fiscal policy proposed by the Ford Administration for the coming year is too restrictive	40	19	41
28. The ceiling on interest paid on time deposits should be removed	76	18	6
29. "Consumer protection" laws generally reduce economic efficiency	24	28	48
30. The economic power of labor unions should be significantly curtailed	32	38	30

TABLE 2—RESPONSE PATTERNS

Proposition	All Mean	All SD	All #	Academic 1 Mean	Academic 1 SD	Academic 1 #	Academic 2 Mean	Academic 2 SD	Academic 2 #	Total Academic Mean	Total Academic SD	Total Academic #	Government Mean	Government SD	Government #	Business Mean	Business SD	Business #
1	2.8	.49	206	2.8	.44	24	2.8	.53	77	2.8	.51	101	2.7	.54	48	2.8	.43	57
2	1.8	.82	204	1.7	.71	23	1.8	.85	78	1.8	.82	101	2.0	.83	46	1.6	.76	57
3	2.2	.86	199	1.6	.82	24	2.3	.85	73	2.1	.88	97	2.2	.82	46	2.4	.84	56
4	2.6	.64	203	2.8	.41	25	2.7	.58	76	2.7	.54	101	2.6	.65	48	2.4	.77	54
5	2.6	.59	208	2.6	.64	25	2.6	.60	78	2.6	.60	103	2.5	.54	48	2.1	.88	56
6	1.7	.75	198	1.2	.52	25	1.9	.80	70	1.7	.79	95	1.8	.71	48	1.6	.73	55
7	2.6	.67	206	2.8	.53	24	2.6	.66	78	2.6	.63	102	2.4	.80	48	2.7	.61	56
8	2.1	.85	201	2.0	.85	23	2.1	.84	75	2.0	.84	98	2.1	.86	47	2.1	.88	56
9	2.6	.65	206	2.8	.56	24	2.5	.66	77	2.6	.65	101	2.6	.58	48	2.5	.71	57
10	2.1	.83	202	2.5	.67	23	2.2	.82	77	2.3	.80	100	2.1	.81	45	1.8	.79	57
11	2.0	.84	198	2.0	.72	22	2.1	.86	76	2.0	.83	98	2.2	.79	44	1.3	.86	56
12	2.3	.73	206	2.5	.66	25	2.5	.70	77	2.5	.69	102	2.3	.71	48	2.1	.78	56
13	1.8	.83	203	1.7	.81	23	1.9	.84	77	1.9	.83	100	1.7	.77	47	1.9	.87	56
14	2.5	.65	205	2.5	.65	25	2.5	.68	76	2.5	.67	101	2.5	.58	47	2.4	.65	57
15	1.3	.59	208	1.4	.71	25	1.4	.59	78	1.4	.62	103	1.4	.61	48	1.2	.52	57
16	2.8	.48	207	2.8	.41	24	2.7	.53	78	2.7	.50	102	2.7	.51	48	2.8	.41	57
17	2.5	.72	202	1.3	.70	24	1.7	.76	77	1.6	.76	101	1.4	.69	46	1.5	.69	55
18	2.3	.77	201	2.6	.49	25	2.5	.70	77	2.5	.66	102	2.0	.87	47	2.2	.79	52
19	1.9	.86	197	2.2	.87	25	1.9	.89	72	2.0	.88	97	1.8	.87	46	1.7	.81	54
20	1.9	.88	200	1.6	.83	24	1.8	.89	76	1.7	.87	100	1.9	.83	46	2.3	.80	54
21	2.1	.76	200	2.0	.75	24	2.2	.75	75	2.2	.75	99	2.1	.78	45	2.1	.76	56
22	2.2	.80	202	2.3	.81	23	2.3	.74	77	2.3	.75	100	2.1	.87	47	2.3	.81	55
23	2.3	.77	201	2.2	.90	23	2.4	.72	76	2.3	.77	99	2.3	.81	46	2.4	.73	56
24	2.3	.77	201	1.0	.20	24	1.5	.82	79	1.4	.77	102	1.4	.75	47	1.3	.65	55
25	2.3	.78	204	2.7	.61	25	2.4	.72	76	2.5	.70	101	2.3	.78	46	2.0	.82	57
26	2.0	.82	200	2.3	.90	24	2.1	.75	75	2.2	.79	99	1.9	.86	47	1.7	.75	54
27	2.0	.90	199	2.3	.92	24	2.0	.92	75	2.1	.93	99	2.1	.86	45	1.7	.84	55
28	2.7	.57	203	2.7	.61	25	2.7	.53	75	2.7	.55	100	2.8	.52	47	2.6	.65	56
29	1.8	.82	203	1.7	.86	24	1.8	.83	78	1.8	.83	102	1.6	.72	46	1.9	.80	57
30	2.0	.79	204	1.8	.78	23	2.0	.80	78	2.0	.79	101	2.0	.76	46	2.1	.80	57

Note: 1 = disagree, 2 = agree with provisions, 3 = agree

Consensus was assessed by calculating an information theory measure, relative entropy (or its compliment, redundancy). This measure of consensus has a range from zero (no consensus) to one (perfect consensus). One interpretation of the measure is that questioning more than one economist would be redundant for questions with relatively high consensus. Conversely, for propositions with low measures, members of the profession can be regarded as entropic or without pattern in their responses.[1]

Table 3 provides the redundancy calculations associated with each proposition in the survey. Since a few individuals refused to answer each proposition within the format provided or did not respond at all, a fourth response category was created, hence four different redundancy calculations are reported.

The scores among the response categories are highly correlated, although different orderings of the propositions result from each measure. The difficulty with the first measure is that only those who "initiated" the fourth alternative generated response in the "other" category. Those who might have done so if it were an option, but who felt constrained to

[1]Relative entropy is defined as actual entropy divided by the maximum possible entropy for the number of outcomes considered, where entropy is the sum of the probability of a particular outcome times the *log* to the base 2 of the probability, i.e., $(-\Sigma p_i log_2 p_i)$. One of the clearest discussions of this measure is found in W. R. Garner We have found "redundancy" to be relatively insensitive to minor differences in proportions and to require really strong contrasts before large measures of redundancy are obtained.

TABLE 3—REDUNDANCY MEASURES

Proposition	(2)	(3)	(4)	(5)	(6)
16	.54	.47	.81	.86	32/1
1	.53	.48	.76	.81	24/1
5	.37	.25	.73	.71	13/1
28	.42	.38	.73	.68	13/1
4	.44	27	.50	60	8/1
14	30	19	.44	.60	7/1
9	.35	.23	.47	56	7/1
7	.37	.25	44	.53	7/1
12	.24	08	20	38	1/3
23	.18	.08	.17	32	3/1
18	.17	.08	.16	.32*	3/1
25	.21	08	.16	.30	2/1
22	.16	.04	10	.24*	2/1
21	.14	.09	04	.22*	2/1
24	.42	.23	44	20**	1/7
15	.43	.38	.60	.15**	1/12
3	.14	.04	04	13	2/1
30	.16	.01	.01	.12	1/1
10	14	.01	.02	.12	1.5/1
8	13	01	.01	10	1.3/1
11	.11	0	0	.08	1/1
26	.12	0	0	06	1/1
17	.26	16	.30	.03**	1/4
2	18	.05	.08	.03	1/2
20	.14	03	.01	02	1/1.3
27	.14	05	01	.02	1/1
19	.11	.02	.01	.02	1/1.3
13	.21	02	.04	.01	1/1.5
29	16	04	08	.00	1/2
6	17	.05	.16	.00	1/3

Note. Column (2) considers four categories. "generally agree," "agree with provisions," "generally disagree," and "other " Column (3) reports the redundancy measures when those who fail to respond within the prescribed format are simply omitted from the sample. Column (4) considers only "generally agree" and "generally disagree" categories and reduces the total sample size accordingly. Column (5) reports redundancy measures under the assumption that all "agree with provisions" responses would fall in the "generally agree" category if only two choices were offered. The ratio of agree to disagree according to the definition of column (4) is also reported in column (6).

answer within the format, are excluded from the category and included elsewhere. Redundancy measures in column (2) reflect that difficulty, since none of them are particularly high. The third column essentially assumes that all those who did not respond within the format provided should not be used to differentiate opinions even though they may have entertained opinions of their own. It also assumes that "agree with provisions" is a distinctly different response than "generally agree."

Columns (4) and (5) in this table are likely the best measures of consensus, depending upon how one wants to consider the differences between "agree" and "agree with provisions." The assumptions utilized for columns (4) and (5) tended to heighten their consensus score, earlier column assumptions pulled their scores down. The fifth column in particular has some interesting redundancy measures. (Scores distinguished by ** are those where a large number of respondents disagree with the proposition, a very small number agree but a

fair number fell in the category "agree with provisions." Scores distinguished by * are those where a large group placed themselves in the "agree with provisions" category and where slightly more chose the "agree" over the "disagree" category.)

Of the thirty propositions, there are twenty with significant consensus at the .01 level: fourteen propositions with which economists agree and six propositions with which economists disagree. Another three propositions (identified by * in the following lists) reach the .05 level of significance. (We used a Kolmorgorov-Smirnov test of the proposition that the middle category, "agree with provisions," falls half way between the "generally agree" and "generally disagree" category. If the middle category is considered to be closer to the "generally agree" category, the number of consensus items would increase. Our conclusions hold under a *chi*-square test also.)

Those issues upon which there is significant consensus and agreement with the proposition as stated include: *1* tariffs; *16* rent controls; *28* interest rate ceilings; *7* minimum wage; *4* cash vs. transfers-in-kind; *9* fiscal policy stimulus; *5* flexible exchange rates; *14* negative income tax; *25* legitimacy of redistributional role; *23* cyclical budget balancing; *18* effluent taxes vs. ceilings; *12* antitrust; *22* regulatory bodies and efficiency; *3* money vs. interest rate targets; *21** money rule is achievable.

Those with significant consensus and disagreement with the stated proposition are: *15* wage-price controls; *24* oil prices and monopoly; *17* pursue money rule; *2* employer of last resort; *29* consumer protection; *6* Galbraith's views; *20** reduce government spending; *13** inflation as a monetary phenomenon. The last proposition (*13*) was however labeled as meaningless by several respondents.

It is clear that economists have reached consensus on many economic issues—a contradiction of the popular image of confusion and disagreement on any issue. However, it would be specious to conclude that we have measured the level of consensus in the profession in general, since we would have to first demonstrate that the thirty propositions were a random selection of all possible economic issues, something we obviously cannot do.

III. Tests of Hypotheses

To test our *ex ante* hypotheses about more agreement on micro-economic propositions than on macro-economic propositions and about more agreement on propositions worded with "can" than on those worded with "should" we created the following matrix of propositions:

Micro "should": *12* antitrust; *22* regulation and efficiency; *28* interest rate ceilings; *29* consumer protection; *30* union power.

Macro "should": *2* employer of last resort; *8* tax indexation; *14* negative income tax; *17* pursue money rule; *19* bond indexation; *23* cyclical budget balancing; *27* Ford's fiscal policy.

Micro "can": *1* tariffs; *4* cash vs. in-kind transfers; *5* flexible exchange rate; *7* minimum wage; *16* rent controls; *18* effluent taxes.

Macro "can": *3* money vs. interest-rate targets; *9* fiscal policy stimulus; *13* inflation as monetary phenomenon; *21* money rule is achievable; *26* Phillips curve.

We excluded proposition *15* wage-price controls, which, although basically micro, has macro implications. The high degree of consensus on this question strengthens the results if grouped with the micro propositions, and only slightly weakens them if it is included with the macro questions.

A 2 x 2 analysis of variance with the redundancy score as the dependent variable results in an F of 9.87 for the micro-macro factor and an F of 6.20 for the can-should factor each with 1 and 19 degrees of freedom. On this basis we reject the hypothesis of micro-macro similarity of response at the .01 level of significance and the should-can similarity of response at the .05 level. The interaction is not significant ($F = 2.921$). The survey seems to support our hypotheses that there is more consensus about micro issues than macro issues and that "can" propositions generate more consensus than "should" ones.

It is clear that those issues which involve interference with the price mechanism and exchange tend to elicit a consensus response. Six of the first seven high consensus items in Table 3 are "price-control" type issues. Of the first ten high consensus items, nine are quite clearly propositions with micro-economic foundations. Proposition *9* (fiscal policy works with less than fully employed resources) is the only intruder, with proposition *17* (the Fed should pursue a money rule) ranking eleventh.

Another way to look at differences in response patterns is to group propositions about similar issues. Consider, for example, those about interferences with the price mechanism, those about fiscal policy, those about monetarist propositions and those about income distribution.

Anyone questioning members of our sample about interferences with the price mechanism would find virtual unanimity of response about their effects (propositions *1*, *7*, *15*, *16*, *28*, and possibly *5*). Moreover, normative judgements matter less here than elsewhere—there was consensus that existing interferences *ought* to be eliminated. In a sense this is supportive of Friedman's hypothesis, value differences being of less importance than differences in what is regarded as theoretically warranted. There was, however, stronger support for the proposition that reducing regulatory power of certain agencies would improve efficiency than for the proposition that consumer protection laws reduce efficiency.

There was consensus that fiscal policy could stimulate the economy and that the budget ought to be balanced over the cycle rather than yearly. There was more diversity of opinion about the appropriateness of the then current fiscal policy and about a possible inflation-unemployment tradeoff.

That the Fed *could* achieve a money rule and that it ought to look at monetary aggregates rather than interest rate targets was generally agreed upon. However, there was also agreement that the Fed *should not* pursue such a rule. There was widespread disagreement about the source of inflation or at least about its source being primarily monetary.

Economists differ little in their opinions that cash is superior to transfers-in-kind and, correspondingly, that the welfare system ought to be restructured along the lines of negative income tax. Redistribution of income was thought to be a legitimate role of government; however, there was less agreement about the current need to redistribute.

Still another way to look at the differences in response patterns is to look for common patterns of response by groups of individuals across all of the propositions. We used factor analysis[2] as a way of examining these common response patterns. The dominant factor, no matter how the sample was cut, was one that indicated correlated agreement with (loadings are in parenthesis): *20* a reduction in government spending (.74); *30* curbs on union economic power (.64); *13* inflation as a monetary phenomenon (.58); *29* consumer protection laws as antithetical to efficiency (.58); *17* the money rule (.58); *22* a reduction in regulatory power as enhancing efficiency (.57); *3* money supply as the important target for monetary policy (.50); and correlated disagreement with: *27* Ford's fiscal policy being too restrictive (.76); *2* the government as an employer of last resort (.72); *10* the necessity of a more equal distribution of income (.63); *25* wage-price controls (.58); *11* reduction in national defense (.57); *6*

[2]Factor analysis suggests the nature of the hypothetical constructs underlying a larger set of empirically obtained variables. As such it is essentially a data reduction and clarification technique. Each factor is characterized by a pattern of loadings for particular empirical variables, the loadings being the correlations with the hypothetical constructs. The procedure determines the fewest possible constructs from the reliable variance in the correlation matrix. The principal components method produces orthogonal factors, each succeeding factor accounting for the maximum amount of variance possible given the pattern of correlations in the correlation matrix. Unless there is a single factor it is usually not parsimonious. The verimax rotation, which retains the criterion of orthogonality, maximizes the clarity of each factor and the purity of each variable's loading on the factor with which it is most associated. This simplifies the interpretation of the factor. An excellent discussion is found in Fred Kerlinger; more detail is available in Harry Harman.

Galbraith (.56); *9* fiscal policy having a stimulative effect (.50). (Combining "agree" and "agree with provisions" does not change the character of the first factor.)

Clearly this factor could be identified with the Chicago School. It shows correlation among politically conservative propositions with a prime focus on monetary policy (particularly the money rule). However, while a nonintervention position may extend to include the money rule, etc., there is no simple connection between these positions and concurrent views about the importance of deregulation, no-income-redistribution policies, or a position on national defense spending. The factor as a whole reflects a political ideology as its dominant theme which only incidentally incorporates a particular theoretical position about the role of money in an economy. It is also interesting that for the total sample curbing the economic power of labor unions is important, but vigorous pursuit of antitrust policy is not (does not appear to be correlated with the listed responses).

The dominant factor under a different (verimax) rotation was also ideological although of a clearly different bent. It indicated consistent correlated agreement with: *6* Galbraith (.77); *24* importance of oil companies monopoly in recent price increases (.70); *11* decrease in defense spending (.61); *2* employer of last resort (.51); *10* distribution of income should be more equal (.43); *15* use of wage-price controls to curb inflation (.41). There were no substantial negative loadings in the factor.

This factor is obviously toward the other end of the ideological spectrum from the first principle component. A surprising aspect of the factor analysis is that it was not capable of explaining much of the variance (18 percent, at most, of the variance could be attributed to either first factor). There are two possible explanations. First, there is simply not much variance in the responses to the micro questions where there was broad consensus across the profession. Hence, this correlated grouping does not appear in the factor analysis. Second, our propositions appear to have hit at enough different values to give very different response patterns individual by individual on macro and particularly on should-type propositions.

IV. Differences Across Occupations

The stratified sample allows for an examination of different response patterns depending upon employment in government, business, general academic or "mentor" academic positions. We were also interested in a more aggregated breakdown which contrasts economists with academic appointments against those in the other two occupations.

Considering first the four employment types, we found that there were significantly different response patterns on only ten of the thirty propositions using a *chi*-square test (the statistic ranges from 20.88 to 13.47, 6 df). It is possible to have group consensus and a significantly different response pattern between employment types, if, for example, the majority of respondents in each strata agree, say, but one group is almost unanimous while the others have more variance in their responses. The more interesting propositions are those in which the differences in response patterns emerge because of patterns of consensus among an employment group that are opposite patterns of consensus among another group(s).

Of the ten propositions with different patterns by employment the two most divergent are for propositions about macroeconomic policy (*26*; *20*). These are followed by the propositions concerning the distribution of income in the economy (*25*; *10*). An appropriate approach to pollution (*18*), Galbraith's description of the *U.S.* economy (*6*), and the role of the oil companies in oil price increase (*24*) also generate differences. The other propositions (*3*; *37*; *9*) are all about macro-economic issues.

The following is a general summary of the differences.[3] When there is consensus on a proposition it is much stronger for the sample

[3]A more detailed statistical breakdown is available from the authors upon request.

of full professors at the seven leading graduate schools than for any other sample. For example, on proposition *24* (concerning the oil price rise and oil company monopoly), 96 percent of the full professors disagreed with the proposition, as contrasted with 65 percent of other academic economists and slightly larger percentages for business and government economists. The differences between the strata, however, are more than simply a degree of consensus. Consider for example, the interesting differences between the two academic subsamples: A majority of the full professors disagree with the "Galbraithian Corporate State" view (80 percent) while the random sample of other academics finds a majority in agreement (60 percent). A similar split also happens for proposition *3* (about the relative importance of money and interest rate targets): 58 percent of the full professor mentor sample disagree with money as the prime target while 74 percent of the sample of other academics believe it to be the appropriate target. Examining other propositions one might conclude that the broader academic sample (younger?) included more people who were ideological, in the sense that the sample tended to be more "Chicago" and more "Galbraithian" and hence, more polarized into subgroups around ideologies.

For propositions *25* (redistribution is legitimate), *20* (reduce government spending), *10* (income redistribution), *3* (money vs. interest rate targets) and *27* (Ford's fiscal policy), academic economists reveal opinions of a distinctly different pattern from those employed in business. Somewhat surprising is the 50–50 split by business economists on the Galbraithian Corporate State view (remember, this view was strongly rejected by the mentor sample and accepted by a majority of those drawn from a larger academic sample). Put differently, splitting the sample into academic and nonacademic employments yields the following differences. The academic group is more in agreement with the notion that unemployment can be reduced by increasing inflation (*chi*-square 13.05 with 2 df), more prone to agree with the notion that the distribution of income in the United States should be more equal (*chi*-square of 10.28 with 2 df), and with the idea that government's role in income redistribution is legitimate (*chi*-square of 9.85 with 2 df). The nonacademic group, on the other hand, agrees more with the idea of reducing government spending (*chi*-square 12.28 with 2 df) and less with the idea of vigorous use of antitrust laws being necessary to reduce monopoly. Consistent with these positions, the nonacademic group rejects the notion that the Ford fiscal proposals were too restrictive.

It should be emphasized that there was no significant divergence of opinions between any strata on questions of a micro-economic nature. Indeed, given the nature of some of the propositions (for example, propositions about price supports), it is interesting how few interstrata differences there are.

V. Conclusion

Consensus tends to center on micro-economic issues involving the price mechanism while the major areas of disagreement involve macro-economic and normative issues. The normative nature of many issues also allows ideological considerations to become important. However, it is clear from this analysis that the perceptions of widespread disagreement are simply wrong. On the other hand, it is true that for many outside the profession the questions of greatest interest are also those that generate the most disagreement within the profession. Hence a good deal of the sampling of economists' advice, which is in turn communicated to the public, comes from the weakest cell in our analysis—macro-economic policy. The problems in this cell are undoubtedly exacerbated since many of the policy statements also involve forecasting of one sort or another. Put differently, the intersection of the greatest interest by the public and hence by journalists with what the profession "knows" occurs in the weakest cell.

In part, this returns us full circle. The difficulty lies, to a degree, not in what there is to communicate, but what is actually communicated and how it is communicated. Indeed, part of the problem is that we have not been successful in communicating what our weak-

est cell is. This, of course, is affected in an important way by the interaction between economists and journalists.

REFERENCES

E. Dale, *New York Times*, Nov. 7, 1971.

Milton Friedman, *Dollars and Deficits*, Englewood Cliffs 1968, 1–16.

W. R. Garner, *Uncertainty and Structure as Psychological Concepts*, New York 1962.

Harry H. Harman, *Modern Factor Analysis*, Chicago 1960.

A. Johnson, "The Economist in a World of Transition," *Amer. Econ. Rev.*, Mar. 1937, *27*, 1–3.

Fred Kerlinger, *Foundations of Behavioral Research*, 2d ed., New York 1973.

Paul A. Samuelson, *Collected Scientific Papers*, Vol. II, Cambridge, Mass. 1966, 1628–30.

"The Furniture Movers," *Business Week*, Jan. 16, 1978, 120.

"Letters to the Editor," *London Times*, Oct. 22, 1971, 13.

———, *London Times*, Oct. 25, 1971, 13.

Who's in the Labor Force: A Simple Counting Problem?

By ARVIL V. ADAMS*

The achievement of full employment or, as it is sometimes presented, the minimization of unemployment has been a major goal of public policy since the economic cataclysm of the 1930's. This goal reflects the implicit belief among policymakers that achieving full employment is the appropriate concern of a manpower policy responsive to the needs of individuals and society. This perception, along with the labor force concepts used to measure progress toward the full-employment objective, has its origin in the surroundings of the depression era and the Keynesian revolution. The events of this period, marked by mass unemployment and related economic hardship, and their conception in economic theory continue to shape contemporary economic policies and labor force concepts.

What proved to be an adequate measure for one set of perceived problems may prove to be inadequate for another, necessitating a change in concepts or methods of measurement. In particular, the relevance of depression era policies and labor force concepts to the present is a question of major importance. Current surroundings have changed, with the growth of income transfer programs and multiple earner families weakening the link between unemployment and economic hardship. As the surroundings have changed, so has economic theory. Led by the resurgence of neoclassical theory and the development of neo-Marxist theories of segmentation, the perception of unemployment and its causes has changed over time.

This paper traces the evolution of economic theory and events, and their impact upon labor force concepts. The relationship of current concepts of employment and unemployment to Keynesian theory and events of the depression era is described and the implications of post-Keynesian theories for these concepts explored. The argument is advanced that current labor force concepts lag behind contemporary economic theories and events. Some directions for change are suggested.

I. The Development of Labor Force Concepts

Current labor force concepts have been part of the economic landscape for nearly four decades, assuming what seems to be a state of immutability. In the most recent review of these concepts by the 1962 President's Committee, the changes were restricted to sharpening and clarifying the gray areas in labor force status. The concept of employment continues to focus on the stock of persons working for pay or profit and among the unemployed, those currently available and actively seeking work.

Yet, when viewed historically, these concepts can be seen as part of a continuum that reflects the influence of changes in economic theory and events. This view is expressed in the early writing of Clarence Long who argued "it is a basic mistake to assume there is only one concept, definition and statistical measure of unemployment" (p. 1). Support for this position can also be found in a review of labor force concepts in the pre-Keynesian period.

Classical economic theory prior to the Great Depression, as summarized by A.C. Pigou, concluded that if wages were flexible no serious unemployment would persist Recognizing that unemployment of a fric-

*Professor of economics, University of Utah, on leave as executive director, National Commission on Employment and Unemployment Statistics. I alone am responsible for the views expressed herein.

tional nature would occur and clearly had done so in several periods of economic crisis during the nineteenth century, classical theory saw the problem as a minor dislocation to be remedied in the long term by wage and price flexibility. The labor force concepts and issues in public policy that emerged from this perception overlooked unemployment and focused instead on employment.

The "gainful worker" concept, as an example, reflected the realities of this period and their conception in economic theory. In a self-regulating market economy where work was usually synonymous with the meeting of needs, interest centered upon the stock of persons in the paid economy. As measured in the decennial censuses from 1870 to 1930, the gainful worker approach counted persons 10 years of age and older by their usual occupation. No effort was made to measure unemployment and new entrants to the job market were excluded from the work force because they were not yet in a gainful occupation.

Given this perception of how the economy worked and the tremendous economic development of the nation in the latter half of the nineteenth and early twentieth centuries, there was little interest expressed in a measure of unemployment. Policy was concerned with such issues as immigration to fill new jobs being created and the threat of monopolies represented by big business and unions. This view of the world was soon to disappear, however, with the turmoil of the 1930's.

The depression brought with it mass unemployment and related economic hardship for those without jobs. The perception of the economy changed, leading to the Keynesian revolution. Pre-Keynesian business cycle theory had implicitly recognized that large-scale unemployment could occur. Its actual occurrence, however, coinciding with John Maynard Keynes' *General Theory*, opened the way for new policies and labor force concepts.

The Keynesian revolution formed the intellectual foundation of current labor force concepts. The events of this period, as explained by Keynes, demonstrated that persistent unemployment could occur. The determinants changed from wage flexibility to those mechanisms controlling aggregate demand. This view of large-scale involuntary unemployment's existence split the classical model down the middle, separating the consumption decision from the labor supply decision. For Keynesians unemployment became a measure of how much effect aggregate economic policy could have on the economy; how much economic slack could be eliminated without any cost to employed members of society.

This perspective underlay the development of labor force concepts in this period. From early experimentation in the 1930's, analysts working on a Works Project Administration program developed a new conceptual framework that was put into practice in the 1940 Census. Both employment and unemployment were measured as stock concepts. A reference period was introduced to measure the stock of unused resources subject to aggregate demand policies at a point in time.

Within this framework, the individual remained as the basic unit of measurement. As a measure of labor supply, an activity concept was used. The presumption was that individuals had to be available and actively searching for work to be counted as unemployed. The use of this criterion avoided measures based on a "state of the mind," and served as an indicator of labor force attachment.

The concepts that emerged from this approach were used to focus on the Keynesian question of involuntary unemployment. Unemployment measures were interpreted as the amount of excess labor supply due to insufficient aggregate demand. Additional data on hours of work assisted in measuring the amount of underutilized labor in the economy. These concepts continue to influence our current labor force measures and their interpretation.

II. Post-Keynesian Theory and Labor Force Concepts

The growth of income transfer programs and increase of multiple earner families following the depression era have weakened

the link between unemployment and economic hardship. The separation of production from consumption has been accompanied by new perceptions of unemployment and its causes. Post-Keynesian economic theories, led by the resurgence of neoclassical theory, have introduced questions about the relevance of current labor force concepts to the present. These questions concern the usefulness of the concepts as guides to full-employment policies, as measures of labor supply, and as indicators of whether individual and social needs are being met in the labor market.

Post-Keynesian theories have added new explanations of unemployment to that of insufficient aggregate demand. Current labor force concepts, however, have been unable to distinguish these explanations and measure their relative importance. As a consequence, the contribution of these concepts to policy debates has been marginal. This can be easily illustrated beginning with the aggregate demand-structuralist debate of the early 1960's. With this debate, emphasis shifted to the measurement of frictional unemployment as a guide to the limits of aggregate demand policies. The need for this distinction followed from recognition that frictional unemployment, unlike the Keynesian concept of involuntary unemployment, cannot be reduced without incurring some cost.

The concepts proved unequal to the task, however, as efforts to measure frictional unemployment remain at the forefront of current debate on full-employment policies. An example of the measurement problems can be found in frictional unemployment attributable to speculative search behavior and unreasonable expectations. Neoclassical search theory stresses the voluntary or speculative nature of unemployment. Some individuals are able to get jobs, but rather than accepting them, hold out for better ones. Present surroundings and the weakened link between unemployment and economic hardship doubtless facilitate this behavior. Conceptually related, the job search of others may be prolonged by unreasonable wage expectations or the lack of qualifications for the jobs sought. Current labor force concepts are unable to distinguish among these forms of unemployment.

The relevance of current labor force concepts to the present has been challenged in other ways by neoclassical theory. The stock concept of current measures reflects a point on the theoretical labor supply function. The interest of neoclassical theory, however, is on the position and shape of this function, and the terms and conditions under which people would change their labor force status. Current concepts are not well suited to this purpose. The rigid application of the activity criterion, for example, precludes counting some of the involuntarily unemployed who have withdrawn from searching. As a consequence, current concepts are ineffective as measures of labor supply. The concepts are of little value to discussions of such issues as the labor supply effects of a guaranteed minimum income or jobs program.

As an indicator of how well individual and social needs are met in the labor market, current labor force concepts continue to focus on the individual. However, neoclassical theory, in reuniting employment and consumption theory, emphasizes the family as a decision-making unit. The majority of individuals are members of a group, most often a family, which uses its labor market capacity to maximize the collective welfare of the group. The extent to which individual and social needs are met in the labor market thereby requires concepts that consider the collective welfare of groups electing to share their resources.

III. Labor Force Concepts and Directions for Change

This brief discussion illustrates the extent to which current labor force concepts lag behind contemporary events and their analysis in economic theory. Stated simply, these concepts, formed in the depression era and the Keynesian revolution, are no longer responsive to the needs of policymakers. The concepts are out of touch with the realities of the present. As part of a continuum in history, changes are needed. The suggestions offered here require greater emphasis on labor force flows, qualitative dimensions of labor force activity, and new measures of labor market-related economic hardship.

The diverse explanations of unemployment

emerging in the post-Keynesian period have shifted attention from the stock concept of unemployment to a flow concept, allowing analysis of the circumstances under which people become unemployed. The flow concept provides insights into movements in and out of the labor force, movements between employment and unemployment, reasons for these movements, and is essential to the measurement of involuntary unemployment and the consequences of employment policies. The shift from a stock to a flow concept does not argue against the classifications of current labor force concepts. It simply recognizes the arbitrary nature of these classifications and the diverse conditions that bring people to a particular classification at a point in time, and the circumstances under which change is likely to occur.

Indeed, this is a subtle distinction, but one which is not unlike that of static and dynamic systems. Although past reviews have left the basic Keynesian stock concept intact, important departures from this concept have appeared in labor turnover data and efforts to measure hidden unemployment. Classification of the unemployed as entrants, reentrants, quits, and layoffs provides an example of attempts to measure the diverse conditions that bring people to a particular classification at a point in time. Representing only a partial answer to the measurement of involuntary unemployment, it nevertheless is an important step toward a flow concept.

The rigid application of the activity concept in turn leads to the classification of some of the involuntarily unemployed outside the labor force. The measurement of so-called discouraged workers, begun in the 1960's, though not considered a measure of hidden unemployment is again a step toward a flow concept. It represents an effort to understand the conditions that bring persons to a particular classification at a point in time and the circumstances under which change would occur.

Keeping in mind the need to develop labor force concepts consistent with current events and explanations of these events, the potential for expanding information on labor force flows is substantial. Current collection procedures allow measurement of the labor force status of persons in a given household periodically over time. At a simple descriptive level, the use of these data to describe gross flows among labor force classifications, although complicated by problems of measurement and technology, is a potentially important source of information about the labor market dynamics which underlie employment, unemployment, and labor force patterns. Although earlier reviews have recommended that gross flow data be developed and the feasibility of doing so demonstrated, they have all but been abandoned by the producing agencies.

Information on labor force flows can also be developed through questions in current labor force surveys. Among the employed, for example, we know nothing about the number of new hires in a given month, their personal characteristics, the jobs they hold, or the methods they used to find employment. For the unemployed, little is known about the type of work sought (occupation, wages, hours), the training and experience of the individual, and the intensity of job search. And for those persons classified as not in the labor force, we have only rudimentary measures of labor force attachment. We do not know, for example, their availability for work or the recency of job search, if any, and its intensity. We have no information on the type of work that would be sought (occupation, wages, hours) or preparation in terms of training and experience.

In the limited space provided, this is only a sketch of the information needed to broaden current labor force concepts to bring them into touch with present realities and to increase their usefulness as guides to full-employment policies and as measures of labor supply. Some of this detail, of course, would be more difficult to obtain than other, due to its subjective nature, but it is needed to understand the conditions that bring persons to a particular classification at a point in time and the circumstances under which change is likely to occur.

Further, labor force status—employed or unemployed—is an inadequate measure of whether individual needs are being met in the labor market. The broad dichotomy of these measures fails to capture the qualitative dimensions of the labor market experience. A

measure of the adequacy of this experience is needed that encompasses those with inadequate employment and earnings and those without jobs who want them. Full-employment policies that focus solely on the dual state of being with or without a job in the Keynesian mold will fall short of their objective of meeting individual and social needs as expressed through the concerns of manpower policy.

Whether or not the adequacy of employment for any individual results in economic hardship depends upon the individual's relation to a family unit and the economic well-being of this unit. The Keynesian concept of unemployment in the depression era could be equated with economic hardship, but inadequate employment today cannot be treated in this fashion, although it may be tied to other forms of hardship, chiefly psychological. In present surroundings, consistent with neoclassical theory, the well-being of the consumption unit must be considered alongside the adequacy of employment for the individual. Milton Friedman, in his criticism of aggregate unemployment as a numerical goal for economic policy, suggests that the size of the unemployment goal is unimportant as long as the unemployed are not suffering and are being retrained for more productive work. This view supports the needs for a family-based measure of labor market-related economic hardship.

REFERENCES

M. Friedman, *Wall Street Journal*, Feb. 3, 1972.

John Maynard Keynes, *The General Theory of Employment, Interest, and Money*, New York 1939.

C. Long, "The Concept of Unemployment," *Quart. J. Econ.*, Nov. 1942, *57*, 1–30.

A. C. Pigou, *Theory of Unemployment*, London 1933.

President's Committee to Appraise Employment and Unemployment Statistics, *Measuring Employment and Unemployment*, Washington 1962, 9–29.

Measuring Economic Hardship in the Labor Market

By DIANE WERNEKE*

The major goal of public policy is to increase the well-being of the nation's population With respect to economic policy, this objective is generally associated with a rising standard of living and the alleviation of poverty. In designing programs directed at economic hardship, public policy has placed priority on self-support rather than economic dependency, but has distinguished between those who are able to work and those who are not. Transfer payments are designed to assist those who are unable to work, while those who can work are assisted by employment and training programs aimed at improving the marketability of their labor.

Given that public policy has taken a two-way approach to the alleviation of poverty, there is a need for corresponding indicators which show progress toward this goal. The measures most commonly associated with well-being or economic hardship are the government poverty line and the unemployment rate. However, neither of these indicators adequately distinguishes whether hardship is due to the failure of the transfer payment system or to inadequate job opportunities. The poverty line provides a global measure of the population in poverty, but tells us little about what is the most effective way to raise the standard of living of this population. While the unemployment rate tracks the overall performance of the labor market, it does not directly correspond to the percentage of those who are able to work and who are experiencing economic hardship. Many of the unemployed are above the poverty line. Rising earnings and the increased availability of alternative and supplementary sources of income have allowed many workers a greater choice in the type of work and pay that are acceptable to them, to leave unsatisfactory jobs, and to weather longer periods of job search. Similarly, the growth in the number of households that have more than one earner has meant that many households may no longer experience total loss of income when one of the wage earners becomes unemployed. On the other hand, there are many persons who are employed but whose earnings do not provide them with an adequate standard of living.

While recognizing that the incidence of economic hardship is high among those in the population who, for whatever reason, cannot work, the focus of this paper is on those who choose to meet their income needs by participation in the labor market, but who are unable to attain a satisfactory standard of living through work. The first part of this paper discusses a concept of economic hardship and some of the issues that arise in its development; the second part suggests one way in which hardship can be measured and illustrates this method by specifying an indicator of economic hardship in the labor market.

I. What Is Economic Hardship?

In the broadest sense economic hardship denotes the failure to achieve or maintain a certain economic status or level of living corresponding to personal expectations. However, in a policy-oriented context, economic hardship is more frequently associated with an absolute income level that is inadequate to provide minimum consumption requirements or basic material needs.

In the absence of assets or transfer payments, a worker's level of living is dependent on two factors: the availability of employment and the adequacy of earnings. Thus, to develop a concept of economic hardship among those who are in the labor market requires simultaneous consideration of the individual's status in the labor force together

*Staff economist, National Commission on Employment and Unemployment Statistics. I alone am responsible for the views expressed.

with the adequacy of the income he or she derives from that status.

Numerous research studies have identified these two conceptual parameters of economic hardship in the labor market and, in translating these into a working definition of labor market hardship, a number of important issues have been raised.[1] Foremost among these issues is, what is meant by an adequate standard of living. Clearly, any choice of an income standard involves a value judgment. Some analysts in this area have used the legal minimum wage as a basis for determining an acceptable level of living, but the most commonly used standard has been the government's poverty line. Though its shortcomings are well known (see, for example, HEW's *The Measure of Poverty*), it represents the official consensus of an absolute level of income adequacy and is probably the best alternative now available, consistent with public understanding.

A second issue raised in developing a definition of labor market hardship is what groups should be counted in a hardship measure. Because such a measure seeks to delineate hardship among those in the labor market, some criteria to determine a degree of labor market attachment are required. Earlier efforts to measure labor market hardship did not, however, develop explicit measures of attachment to the labor market; rather, most eliminated categories of workers such as students, retired persons, or in some cases, all so-called secondary workers, from the hardship count. The rationale for these exclusions was generally to remove groups whose labor market attachment was perceived to be tenuous, such as full-time students whose studies may have dictated work patterns.

However, there are several problems in defining labor market attachment by demographic characteristics. Foremost is the possibility of excluding students or elderly persons whose labor market problems result in economic deprivation. Also, the exclusion of secondary workers tends to bias a measure against women. Because the classification of a primary worker is based on the highest earner in the family and women have not generally attained equal opportunity in the labor market, excluding secondary workers from a hardship count will exclude most women in husband-wife families from the population base.

A third issue raised in defining labor market hardship is whether such a measure should reflect only the lack of individual opportunities for finding suitable and productive employment or, in addition, the lack of family income resulting from inadequate employment. Current labor market statistics focus on individual status: a person is employed, unemployed, or not in the labor force. However, economic well-being or hardship is more commonly thought of in terms of a family or household whose members share in the consumption of available resources. Thus, in defining labor market hardship, an issue is whether to focus on labor market problems which result in economic deprivation for the household or for the individual, regardless of family situation. Most analysts have been concerned with hardship in a family or household context and have thus viewed an individual's earning inadequacy together with that individual's relationship to a household. Thus, in addition to applying an income standard to individual earnings, these analysts also apply a family income adequacy cutoff to indicate the number of people working or seeking work who do not have other working family members or other sources of income which ameliorate the consequences of their own labor market problems. However, one problem in applying a measurement of economic hardship to a family or household unit is the underlying assumption of a fixed family configuration. A family with an adequate total income may contain an individual who would prefer to live outside the household but cannot because of insufficient resources. Thus, using a household base may understate economic hardship to the extent that it does not include those who would prefer different living arrangements but cannot make them with available resources.

[1] Though not exhaustive, a list of authors would include William Spring, Bennett Harrison and Thomas Vietorisz; Herman Miller; Sar Levitan and Robert Taggart; Vietorisz, R. Mier and Jean-Ellen Giblin

A related issue that arises in this context concerns the period of time over which to view an individual's participation in the labor market and the income derived from participation. Labor force status is currently thought of in terms of a survey reference week or month, but this time frame may not be an appropriate horizon for revealing the dominant labor market situation of an individual whose status viewed at any one point in time may be a temporary circumstance. Similarly, a question arises as to the time horizon for income measurement for the purposes of comparison with an income standard. Since many of the unemployed experience relatively short periods of joblessness, the use of a short period over which to measure income may reflect temporary circumstances, while others might receive high seasonal earnings and yet low annual income.

To define economic hardship in terms of the adequacy of earnings or income and the availability of employment opportunities requires the resolution of the issues discussed in this section. This, of course, involves value judgments on the part of the analyst, and arbitrary assumptions must be made at each stage of developing a working definition.

II. A Further Effort to Define and Measure Hardship

At the outset it might be helpful to set out the basic conceptual and technical criteria used here to determine earnings, income and employment adequacy. A worker is counted in the measure of labor market hardship if: 1) he or she earns less than the poverty threshold during the calendar year, and also is a member of a household whose total income resources are less than twice the poverty standard; and 2) he or she has demonstrated a significant amount of labor market activity either by working, seeking work, or some combination of the two for at least forty weeks during the same calendar year.

Two exceptions to this labor force attachment test are adopted for this measure: 1) discouraged workers, who did not work at all during the year but who reported their inability to find work was the reason for not working, are included if they also demonstrated a reasonable attachment to the labor force by spending at least fifteen weeks during the calendar year looking for work; 2) individuals who voluntarily chose to work part-time for most of the year were excluded on the rationale that their circumstances were not sufficiently the result of labor market problems, but rather the effect of other factors affecting their choice of part-time employment.

This specification was applied to data from the March supplement to the *Current Population Survey* (*CPS*) reporting the work experience of the population for the nation in the preceding calendar year. This annual supplement is readily available since 1967 and includes information on the labor market activity and income of all household members. The survey reports earnings separately from income. It also contains the demographic characteristics for each member.

A. *Derivation*

Altogether there were about 108 million persons in the labor force for all or part of 1976. In addition, there were about 700 thousand persons who were discouraged and did not work at all but who had sought work at some time during the year. There were about 9 million persons working part-time voluntarily for more than half of the year. Including discouraged workers and excluding voluntary part-time workers brings the total to about 99 million people to which the earnings, income, and labor market attachment criteria were applied to derive the hardship measure.

The mechanics of constructing a hardship measure can be illustrated as follows: the 99 million labor force is first used to derive the number who had been in the labor market forty weeks or more (about 74.9 million). This figure is then increased by those who were discouraged but had previously sought work for at least fifteen weeks in that year (yielding 75.6 million). Applying the earnings criteria, there are about 15.5 million below the poverty line. Applying the family income criteria to this group results in about 7.6 million workers

TABLE 1—LABOR MARKET HARDSHIP INDEX BY DEMOGRAPHIC CHARACTERISTICS, 1976

Characteristic	Percent Experiencing Hardship
Total	10.0
Men	8.7
Women	12.2
Blacks	24.2
Hispanics	20.7
Whites	8.3
Other minorities	14.4
Age	
16–19	26.4
20–24	12 9
25–64	8.7
65 and over	15.8
Husband-wife family	8.4
Male head	11.8
Female head	24.2
Unrelated individuals	10.7

Source: Special tabulations by the Bureau of Labor Statistics for the National Commission on Employment and Unemployment Statistics

both earning less than the poverty line and in a family receiving a total family income of less than twice the poverty line in 1976. This compares with some 20 million people who experienced some unemployment during that year. The hardship group represents 10 percent of all those in the labor market forty weeks or more or discouraged (but sought work for at least fifteen weeks) in 1976.

Low earnings rather than unemployment or discouragement appeared to be the predominant factor contributing to this measure of hardship as 51 percent of those in the hardship group had worked for all forty weeks, 40 percent had experienced at least one week of unemployment, and only 9 percent were discouraged workers. Indexes were also derived for sex, race, age, and family status. The incidence of employment and earnings inadequacy varied significantly among different groups in the population. Generally, women and minorities were more likely to experience hardship than other groups. Workers in husband-wife families were less likely to be counted while a disproportionately large number of workers in female-headed families and unrelated individuals fell into the hardship group. These results are shown in Table 1.

B. *Secular and Cyclical Changes*

The hardship measure was calculated for calendar years 1967–76, the latest year for which the data are available. This period encompassed substantial variations in labor market conditions. Tight labor markets prevailed between 1967 and 1969 in contrast to that experienced during most of the succeeding years. In 1970 unemployment rose substantially and it remained high throughout 1971. The recovery in the labor market which got underway in late 1972 was short-lived, however, as the 1973–74 recession pushed unemployment to a postwar high. By the end of 1976, the jobless rate was still above 7 percent, although employment had been growing briskly throughout the year.

Over this period, the hardship index moved in the same direction as the unemployment rate, rising as conditions deteriorated and declining again with improving labor markets. However, two important differences in the behavior of the unemployment rate and the hardship index deserve mention. Secularly, the unemployment rate showed an upward trend over the period while the hardship index did not. Secondly, although the cyclical movement of the hardship index was similar in direction to that of the unemployment rate, the hardship index showed much smaller oscillations.

The reasons suggested by analysts for the secular rise in the unemployment rate include the changes in the demographic composition of the work force, the increased availability of income transfer payments, and the rise in the number of multiearner families. These factors apparently did not have the same effect on the hardship index. In fact, the rise in multiearner families and the increase in transfer payments may have alleviated economic deprivation and reduced the number included in the hardship measure.

Cyclically, the magnitude of the changes in the hardship index was less than the unemployment rate. The explanation for the difference in the amplitude of fluctuations is that

unemployment is only one component of a hardship total and hence any percentage change in this group directly alters a hardship index to a lesser extent. Also, only those unemployed with inadequate incomes are counted as experiencing hardship and many who were forced into idleness when labor market conditions deteriorated were members of households with incomes exceeding a minimum adequate standard of living. Thus, although the hardship measure increased during a recessionary period, it did not rise proportionately as much as overall unemployment measured by the *CPS*. In comparison to the poverty index, the hardship index varied to a greater degree in both the downturns and recoveries in economic activity during the 1970's.

Focusing on the demographic trends of workers experiencing hardship, the data suggest that women made no real progress in narrowing the differential relative to men over the 1967–76 period. However, there is evidence that the situation of black workers relative to their white counterparts improved over the period as the downward trend of the index for black workers did result in a noticeable narrowing of the black-white differential. The position of those of Spanish origin was roughly unchanged over the period for which there were data available on this group.

The foregoing discussion of the level and trends in an economic hardship measure suggests that such a measure could be a useful complement to our present labor market data system. The hardship measure shows that the basic needs of many labor force participants have not been successfully met by the labor market. Many are unemployed or discouraged, but more have been working throughout the year. A hardship measure also shows that while recession intensified hardship conditions, the problems are not rooted primarily in the business cycle, but rather are continuing structural problems.

REFERENCES

Sar A. Levitan and Robert Taggart III, *Employment and Earnings Inadequacy: A New Social Indicator*, Baltimore 1974.

H. P. Miller, "Subemployment in Poverty Areas of Large U.S. Cities," *Mon. Labor Rev.*, Oct. 1973, *96*, 10–17.

W. Spring, B. Harrison, and T. Vietorisz, "Crisis of the Unemployed," *New York Times Mag.*, Nov. 5, 1972.

T. Vietorisz, R. Mier, and J. Giblin, "Subemployment: Exclusion and Inadequacy Indexes," *Mon. Lab. Rev.*, May 1975, *98*, 3–12.

U.S. Department of Health, Education, and Welfare, *The Measure of Poverty*, Washington 1976.

U.S. Department of Labor, *Manpower Report of the President*, Washington 1967.

U.S. Bureau of the Census, *Current Population Survey*, Mar. 1977 Suppl.

Counting the Labor Force with the *Current Population Survey*

By CURTIS GILROY*

How convenient it would be to prepare unbiased estimates of the labor force from the *Current Population Survey* (*CPS*) by simply multiplying the raw data tabulated by questionnaires by the reciprocal of the sampling ratio (about 1,300). Since the *CPS* is designed to allow each housing unit in the population an equal chance of being selected, it should be theoretically possible to prepare such estimates this way. Unfortunately, sampling is rarely this simple, and certain adjustments must be made to enhance the accuracy of the sample data.

This is not to say that the *CPS* does not have a reasonable sample design to provide accurate national estimates (see U.S. Bureau of Census; Philip McCarthy). However, as with any sample survey, there is the possibility that the sample is not perfectly representative of the total population, and that all units designated for inclusion in the sample may not be interviewed.

I. Accuracy of *CPS* Data

The accuracy of an estimate from the *CPS* is measured by the difference between the estimate and the true value in the population. The ultimate goal is to design and implement the survey in such a way that this difference—or total error—is minimized.

The total error of a sample estimate is composed of two elements: imprecision and bias. Imprecision refers to the fact that, if a sample were replicated many times, the estimates from the samples would differ from each other, even though the same sampling method, interviewing procedures, and questionnaire design were used.

This variability among the sample estimates would occur because different people were surveyed in the different samples and because even the same person may provide different responses in separate replications. The imprecision of a sample estimate is measured by the extent to which the estimate is expected to differ from the average of all estimates resulting from a large number of replications. If the average of these replications equals the true population value, the sample estimate from a single replication is said to be unbiased. In such a case, the total error would be due to imprecision of the sample estimate.

If, however, the average sample estimate differs from the true population value, the estimate is biased. The degree of bias is measured by the size of this difference. Where bias exists, too, the total error is a function of both imprecision and bias.

Although the measure of imprecision in the *CPS* data is relatively simple to determine, it is particularly difficult to estimate the effect of bias on the estimates, because the measurement of bias implies knowledge of true population values which are generally unknown. Sources of bias include inconsistent respondent replies due to differences in interpretation of questions and perceptions of reality by respondents, incorrect recording and coding of information by interviewers, data processing mistakes, imprecise estimation of values for missing data, and failure to include all units in the sample (see Camilla Brooks and Barbara Bailar). Although the standard error is often used as a proxy for the total error of the estimate, it does not measure the bias component of this error. Moreover, where bias is large, the standard error is an unsatisfactory measure of accuracy.

This paper will focus on sources of bias in the *CPS* and will emphasize the importance of measuring its impact in counting the labor

*Staff economist, National Commission on Employment and Unemployment Statistics. I am grateful to Philip McCarthy and Gary Solon for substantive comments on earlier drafts. I alone am responsible for the views expressed herein.

force and in evaluating the accuracy of sample estimates. As a result, the need for published information about the sources of bias has become more pronounced. Hopefully, we can eventually quantify their impact for inclusion in an estimate of the "total" error.

II. Sample Rotation

Like many large-scale repetitive surveys, the *CPS* is designed to interview the same household more than once. Specifically, a household selected for the survey is interviewed for four months, dropped from the survey for eight months, and then interviewed for an additional four months. The household then leaves the sample permanently. Under this 4-8-4 rotation scheme, each month one-eighth of the sample is interviewed for the first time, one-eighth for the second time . . . and one-eighth for the eighth and last time. Seventy-five percent of the sample in a given month is interviewed in the next month too, and 50 percent is interviewed in the same month a year later.

This appears to be a reasonable sample scheme in that selecting an entirely new sample every month would be more costly, while at the other extreme, keeping households in sample longer than that required by the 4-8-4 pattern might impose an excessive burden on respondents. Most important, however, the month-to-month overlap in the samples allows more precise estimates of change between adjacent months, and it is the *change* rather than the *level* in a statistic that is often most useful for analysts and public policymakers. In terms of statistical reliability, the sample design is successful in meeting the objective of month-to-month change in national estimates.

Under the 4-8-4 scheme, each month's sample contains eight rotation groups, each of which comprises a national sample. Therefore, in a particular month it is possible to derive from these groups eight separate estimates of a labor statistic such as the number unemployed. Theoretically, the eight estimates should be the same except for random differences due to the imprecision of sample estimates. But experience has revealed large, systematic differences among the data from different rotation groups. These differences are referred to as "rotation group bias." For example, the estimate of unemployment based on the first rotation group (the one in the sample for the first time) is typically about 10 percent greater than the estimate based on all eight groups.

The direction of rotation group bias is unknown. It has been impossible, for example, to determine which unemployment estimate is accurate—the one from the first rotation group, the one from all eight, or none. Measurement of the bias and development of methods to remove it require an understanding of its causes, about which there has been considerable research but little of a conclusive nature. Rotation group bias is undoubtedly a manifestation of a host of bias sources such as proxy respondents, telephone interviewing, interviewer-respondent relationships, and the like. The use of the longitudinal character of the *CPS* may assist in shedding some light on the nature of rotation group bias as households are followed throughout their stay in the sample. Much research needs to be done in this area.

III. Noninterviews

Loss of coverage from either noninterviews or incomplete interviews can bias estimates since the sample results are no longer derived from a sample which was designed to be representative of the total population. Loss of information from noninterviews is most problematic.

To compensate for noninterviews, the sample estimates are adjusted by assuming the labor force status of noninterviewed households is distributed like that of persons with similar characteristics from interviewed households—the so-called "comparability assumption." This adjustment weights each observation by the reciprocal of the particular interview rate to maintain the overall sampling ratio.

One obvious drawback to this adjustment procedure is that it is limited to geographic and racial factors. The introduction of other demographic characteristics would improve

TABLE 1—NONINTERVIEW RATES BY REASON, SELECTED YEARS 1954–76

Period	Total	No one home	Temporarily Absent	Refusal	Other
1954–55	4.2	1.6	1.6	.6	.4
1962	5.0	1.9	1.8	1.0	.4
1967	4.9	1.4	1.2	1.8	.5
1972	4.0	1.0	1.0	1.8	.3
1973	4.3	1.1	1.0	1.9	.4
1974	4.1	.9	.9	2.0	3
1975	4.2	.9	.8	2.2	.2
1976	4.4	.8	.8	2.6	.2

Source: U.S. Bureau of the Census.

the imputation process. Further, one could either substitute the prior month's labor force characteristics for noninterviews that had been interviewed or could use their detailed demographic characteristics in the imputation process.

The *CPS* has maintained a consistently low noninterview rate over the past twenty years (see Table 1), despite the changing habits and attitudes of a more mobile population. Within the total sample, noninterview rates differ considerably by the month in which each household is in sample. The rate is high in the first month, then falls, is high again in the fifth month after the household has been out of sample for eight months, and then declines once more. This pattern holds for all categories except the refusal rate, which increases almost continuously.

Refusals are particularly troublesome since they have gradually become a substantial portion (60 percent) of all noninterviews. Data show that the longer a household is in the sample, the more likely it is to refuse the interview as respondents feel the burden of continued interviewing. The noninterview rate, moreover, is largest for households participating in income supplements to the *CPS*, regardless of month in sample, indicating that content as well as length of the survey instrument are important elements in increasing response and reducing bias.

IV. Ratio Estimates and the Census Undercount

One of the several special ratio adjustments modifies the sample estimates in a number of age-sex-race groups according to independently derived current estimates of the population. These independent estimates are prepared by inflating the most recent (1970) census figures for these groups to include the estimated census undercount, then adjusting the results for subsequent change flowing from aging, births, deaths, and net migration, and then deflating them to the census level.

Although this method preserves the actual pattern of population change over time in any age group, the portion of the population missed in the census is still not included in the labor force estimates. This creates the problem of increased bias in the estimates, if the omitted population has a different distribution by employment status in each age-sex-race category than the enumerated population, and leads each month to a lower estimate of the labor force than a fully accurate census would reveal.

The Census Bureau's best estimate of the undercount for 1970 was 5.3 million, or 2.5 percent, with coverage loss greatest for black males (9.9 percent) and least for white females (1.4 percent). In all, about 4.1 million persons 16 years of age and over in the civilian noninstitutional population were missed in the 1970 Census.

The Census Bureau has justified its exclusion of the undercounted population in labor force estimates for several reasons: 1) because the characteristics of the undercounted population are unknown, they may not have the same characteristics as the counted population; 2) the estimates of the undercount for particular subgroups are subject to error; 3) adding the undercount would produce a

TABLE 2—EXTENT OF CENSUS UNDERCOUNT AND UNDERREPRESENTATION IN *CPS* SAMPLE OF PERSONS 20–49 YEARS, BY SEX, AGE, AND RACE (Shown in Percent)

	Net census undercount relative to independent estimates				Underrepresentation in *CPS* sample relative to census figures[a]			
	White		Black		White		Black	
Age	Men	Women	Men	Women	Men	Women	Men	Women
20–24	2.5	1.1	8.7	3.4	6.5	5.9	22.4	9.9
25–29	4.7	2.8	15.6	6.6	6.2	2.3	14.4	5.8
30–34	4.0	2.0	14.4	3.7	4.7	1.8	21.4	11.9
35–39	4.1	0.8	17.8	4.6	4.2	2.5	7.9	8.8
40–44	3.2	0.1	16.3	3.5	3.7	1.9	18.1	8.2
45–49	3.5	0.5	13.3	5.0	3.0	2.1	12.6	5.0

Source U.S. Bureau of the Census
[a]Data relate to average for July 1974–June 1975 but are typical of the situation encountered every month.

disturbing discontinuity in many *CPS* series; and 4) no usable estimates of the undercount are available for use in compiling *CPS* data for states and local areas. These objections are discussed in order.

1) A counter to the objection to assuming that the undercounted population has the same or comparable characteristics as their enumerated counterparts is that it is perhaps more objectionable to assume that unenumerated persons do not exist at all. Furthermore, studies have shown that the labor force status of the undercounted population is *not* necessarily different than that of the counted population (see Deborah Klein). Finally, comparability assumptions are already made in compiling the *CPS* data in the noninterview adjustment as noted above.

In two-thirds of the cases in Table 2, the rate of underreporting for some groups in the *CPS* sample, relative to controls derived from the census count, is actually larger in magnitude than the census undercount relative to the independently derived estimates of the true population. This *CPS* underrepresentation has always been offset by assuming that the unreported persons have the same characteristics as enumerated persons in the same age-sex-race group. Thus, adjusting for the census undercount would merely be an extension of present practice—and of a lesser magnitude—to bring the *CPS* estimates in line with the true population figures.

Table 3 shows how much the adjustments currently used to bring the *CPS* sample of persons 20–49 years of age into line with the

TABLE 3—FACTORS APPLIED TO CPS SAMPLE TO CORRECT FOR UNDERREPRESENTATION OF CENSUS AND *CPS* FOR PERSONS 20–49 YEARS, BY SEX, AGE, AND RACE

	Present factors to correct for *CPS* underrepresentation				Factors to correct for census undercount and *CPS* underrepresentation[a]			
	White		Black		White		Black	
Age	Men	Women	Men	Women	Men	Women	Men	Women
20–24	1.07	1.06	1.29	1.11	1.10	1.07	1.41	1.15
25–29	1.07	1.02	1.17	1.06	1.12	1.05	1.39	1.14
30–34	1.05	1.02	1.27	1.14	1.09	1.04	1.48	1.18
35–39	1.04	1.03	1.09	1.10	1.09	1.04	1.33	1.15
40–44	1.04	1.02	1.22	1.09	1.07	1.03	1.46	1.13
45–49	1.03	1.02	1.14	1.05	1.07	1.03	1.31	1.11

Source. Table 2.
[a]Product of *CPS* underrepresentation factor and census undercount factor.

census population controls would have to be increased in order to include the census undercount. A multiplication factor of 1.10 for white males 20–24 years old, for example, is needed to inflate the estimate to the "true" population figure, compared with 1.07 to bring it into line with an incomplete population count.

2) Errors in the undercount itself are quite small at the national level. The accuracy of the estimated undercount depends on the demographic techniques chosen as well as on coverage and reporting errors in the basic demographic data on age, sex, and race that go into the estimate. The vital statistics are reliable, but the data on immigration are inaccurate and make the official population undercount subject to its own underestimate.

The Census Bureau provides three estimates by age, sex, and race which are very similar, although based on different procedures, data, and assumptions. Although each set of estimates is subject to unknown error which cannot be precisely measured, the undercount estimates for most demographic groups display little variation. This lends credence to the Census Bureau's choice of a "preferred" set of estimates—a composite of the three estimates above.

3) Although the introduction of a census undercount adjustment would result in a significant break in *levels* for such series as employment and unemployment, it would have a negligible effect on *rates* and not disrupt their comparability over time. Discontinuity in a time-series is difficult to manage, however, and care would have to be taken to insure that the transition was as smooth as possible. Labor force estimates would be more realistic and the overall unemployment rate—the most widely used indicator—and all other rates would be virtually unchanged.

4) Estimates of the census undercoverage are now available for each state. Because the true population is unknown and the range of error in the estimates therefore impossible to specify, the state undercount estimates must be used with extreme caution. A major problem is that interstate migration statistics are not collected.

Even without reliable state undercount estimates, however, state data could be adjusted according to national undercount estimates. In fact, state estimates are presently prepared by using the national second stage ratio adjustment weights. The question then is not *whether* to adjust area data to national population controls by demographic group; this is already done. The question is *which* control to use: incomplete population figures or figures adjusted for the census undercount.

The implications for inclusion of the undercount in the state estimates are important since much federal funding to states is made on the basis of population. Because the estimates of underenumeration reflect substantial variation between states, the distribution of funds could be significantly altered by use of corrected population figures instead of often outdated census figures.

V. Response Errors

Bias from response errors from completed questionnaires can emanate from respondents, interviewers, or their interaction. Respondents may deliberately provide false information, be simply misinformed, or not know what the correct answer is; the interviewers may ask a question improperly or misunderstand the respondent and record something the respondent did not mean; or the question may be misunderstood. Two principal sources of response error of particular concern are the effect of proxy respondents and the more widespread use of telephone interviewing.

Proxy Respondents. Although the *CPS* is a sample of households, labor force information on all eligible persons in that household is obtained from any responsible person. The interviewer is instructed to interview the most knowledgeable household member, but any adult 14 years old and over is eligible to be the respondent. Obviously, the use of proxy respondents may lead to response errors if the surrogate's answers differ from those that would have been obtained if the household member had actually been interviewed. Although of vital concern for all questions in the *CPS*, it is most important for questions involving perceptions and attitudes.

Telephone Interviewing. Because of the widespread use of telephone interviewing, there is growing concern about its effects on the data. Nearly 60 percent of all interviews were conducted by telephone in 1976, compared to only 45 percent in 1969. Over four-fifths of all interviews during months when telephone interviewing is permitted are now conducted by telephone. There is the presumption that personal interviews would provide more accurate information than telephone interviews, but no definitive quantitative analysis can verify this. More testing on this as well as the effect of proxy respondents is needed.

VI. Conclusion

The *CPS* is the cornerstone of our labor market information system, the only data base providing monthly statistics on the labor force, and the world's best documented and most imitated survey. The Bureau of the Census prepares standard errors as measures of imprecision of the estimates. Although the Census Bureau has catalogued the various sources of bias in the survey, it has not included measurements of these in the error estimates. It is important that users of the data be made more aware of the nature of biases in the estimates and that the Bureau of the Census proceed in the direction of eventually providing an approximation of the total error of an estimate—including imprecision and bias—which will tell us more about how accurate is the count of the labor force from the *CPS*.

REFERENCES

C. A. Brooks and B. A. Bailar, "Nonsampling Errors in the Current Population Survey as They Affect the Employment Statistics," unpublished paper, Bureau of the Census, Jan. 1978.

D. P. Klein, "Determining the Labor Force Status of Men Missed in the Census," *Mon. Labor Rev.*, Mar. 1970, *93*, 26–32.

P. J. McCarthy, "Methodological Issues in the Household Survey," unpublished paper prepared for the National Commission on Employment and Unemployment Statistics, 1978.

U.S. Bureau of the Census, "The Current Population Survey: Design and Methodology," tech. paper no. 40, 1978.

MACROECONOMICS: AN APPRAISAL OF THE NON-MARKET-CLEARING PARADIGM

Second Thoughts on Keynesian Economics

By ROBERT J. BARRO*

My view in the early 1970's of Keynesian, non-market-clearing-type models was that the soundness of their theoretical structure hinged on an as yet absent theory of the stickiness of wages or prices. The application of contracting theory to macro analysis seemed promising in this respect. The presence of employee risk aversion or of transaction costs associated with market arrangements—which could include elements of capital that were specific to employment or other aspects of production and exchange—seemed to motivate some long-term, implicit or explicit agreements about wages or prices. In particular, a sluggish adjustment of wages to current economic conditions could be rationalized by this approach.

Further consideration of the contracting model suggests that its rationale for sticky wages and prices—as far as it goes—does not explain the key features of Keynesian analysis with regard to the determination of employment and output. For example, long-term labor agreements do not imply a failure of employment to increase when all parties to the agreements perceive that they could be made better off by such a change. The so-called involuntary unemployment of Keynesian models—that is, a situation where everyone perceives accurately that the marginal product of labor exceeds the marginal value that potential workers place on their time—is not compatible with efficient labor agreements. Even in contracts that specify, *ex ante*, the value of nominal wages over some interval of time, it would be mutually advantageous for workers and firms to determine levels of employment in an efficient manner. The contracting approach may rationalize some departures of real wages from the marginal product of labor and/or the marginal value of worker time, but it does not imply that levels of employment would differ significantly from the (efficient) values that would have been attained under flexible wages. Rather than rationalizing the non-market-clearing model as a useful "as if" approach, contracting analysis suggests that—despite the possible existence of "sticky" wages—the continuous market-clearing model may provide a satisfactory framework for the analysis of employment and output. Notably, the approach suggests that such market features as sticky wages or the apparent non-price, quantity rationing associated with layoffs would be of secondary interest in analyses of business cycles. Since the prevailing wage need not represent the marginal product of labor, the presence of "excess labor supply" at this wage need not signal involuntary unemployment in any economic sense.

The conclusions derived from the contracting model can be generalized by observing that the key assumption of Keynesian analysis is the inefficiency of some aspects of private sector activity in comparison to corresponding activities carried out by the government. This central feature is, of course, the underlying basis for the policy activism that typifies Keynesian thinking. In some simple "disequilibrium" macro models, relative private sector inefficiency is represented by sticky wages or prices, in contrast to the flexibility of such government policy instruments as the money supply, taxes, or expenditures. Technical limitations of the private market in the coordination of production and exchange—as reflected in wage-price stickiness and the associated determination of employment and output through a non-price rationing process—are remedied through the superior coordinating

*University of Rochester. I have benefited from comments by Herschel Grossman, Bob Hall, and Ben McCallum. The National Science Foundation has supported this research.

skill of the government—as embodied in these models in the judicious use of monetary and fiscal policy or in the appropriate direct regulation of prices and quantities.

Presumably, sticky wages or prices are not intended to be taken literally as the source of private sector inefficiency. The underlying problem must reflect some deeper economic elements, such as imperfect information about the present or future, factor mobility costs, or some types of significant transaction costs. Although some of these elements are probably important in business cycle analyses, it is not apparent that they imply relative efficiency of the government over the private sector in handling such economic disturbances as oil crises, harvest failures, or even autonomous changes in liquidity preference or the perceived marginal product of capital—if such shifts occur on a significant scale. For example, uncertainty and mobility costs seem to imply that the allocation of resources is a difficult problem—not that the government can assist in allocation through active use of its macro-policy instruments. In any event the theoretical case for activism would, as in areas like industrial organization and the production of "public goods," require as a first step some serious analysis of private market "failure." For example, a frequently suggested macro-policy response to the oil crisis involved expansion of the money stock or the government deficit. I do not see how to construct an economic analysis of private sector allocation that would imply that the necessary and difficult private adjustments to this type of real—unexpected, but presumably perceived—disturbance would be assisted by an increase in the quantity of money. (I also do not see how this policy response would be called for on grounds of income distribution.) The observation that a monetary expansion might be helpful in models where some prices are arbitrarily held fixed does not seem illuminating.

It is not difficult to construct a theoretical model in which the natural rate of private output is too low, relative to some ideal, because of external effects. For example, the taxation of market earnings and the existence of welfare programs for the unemployed drive a wedge between private and social product, which would imply "insufficient" output on average. This observation would seem important for the design of tax and welfare programs. However, these types of external effects produced by government intervention (which may or may not be warranted on other grounds) do not have obvious relevance for the business cycle or for the usual forms of macro-stabilization policies.

A typical feature of macro analysis is that government intervention into the economy is recommended without bothering to describe the supposed externality or private market failure that underlies the call for policy activism. As a recent example of this tendency, consider the proposal for a tax-based incomes policy (*TIP*), which involves a tax penalty for price or wage changes above some amount and a reward for changes below this amount. I honestly have no idea what sort of private market failure or externality is supposed to rationalize this sort of government interference with the price-setting process. (However, I'm sure it has nothing to do with the triangle under the money demand function that is occasionally used to measure the welfare loss from anticipated inflation.) It is unclear why there is some asymmetry that leads individuals or firms toward "excessive," rather than insufficient, price changes. Casual arguments about external effects from "price leadership" or the like do not seem helpful in this respect. Additionally it is unclear whether the *TIP* proposal is directed at the costs of the average rate of inflation, the uncertainty of inflation, or to some perceived interplay between (I assume, unanticipated) inflation and unemployment. It is also hard to reconcile the plan with the irresistable link between monetary expansion and inflation, although the proponents may have in mind some subtle pressure on the money supply process. In any case the theoretical rationale that has been presented to support the *TIP* plan or other forms of general price controls has not been on the same level of economic analysis as that—weak as it may be—which has been provided to defend government regulation of certain industries, control of pollution, and so on. This lack of a theoretical argument might be provisionally acceptable in the light of supporting empirical evidence on the benefits

of such a policy but the record of previous price control programs does not seem impressive.

Modern courses in macroeconomics utilize price theory to a considerable extent. However, these "micro foundations" are usually limited to the formulation of sectoral supply and demand functions, rather than to the analysis of "general equilibrium." Despite oddities in some earlier treatments of labor supply, the serious problem with non-market-clearing-type models are not in the characterization of supply and demand, but rather in the neglect of the other branch of price theory: namely, supply equals demand. Supply not equal to demand as a basis for quantity determination in non-market-clearing models is not on the same analytical level as supply equals demand. The latter mechanism implies that—at least in a direct sense—the private market manages to exhaust trades that are to the perceived mutual advantage of the exchanging parties. On the other hand, by mechanically leaving opportunities for mutually desirable trades, the non-market-clearing approach makes government policy activism much too easy to justify. When the arbitrariness of supply unequal to demand is replaced by a serious explanation, such as imperfect information about exchange opportunities, for the failure of private markets to achieve some standard of efficency, the case for government intervention becomes much less obvious.

Let me now consider some specific issues relating to information and macro policy. I begin with the role of expectations in macro models. Thanks especially to the work of Robert Lucas, we have a much better idea of the significance of well-informed private expectations in macro analysis. This significance arises in at least three areas: positive analyses of the effects of monetary and other shocks on economic activity; analyses of the role of government polices; and evaluations and carrying out of econometric estimation. Nonetheless, I agree with the view that rational vs. nonrational expectations is not per se the key division between Keynesian and non-Keynesian models and, accordingly, is not the essential basis for a division between activist and nonactivist policy conclusions. The formation of expectations is one dimension in which the private sector might operate less (or even more?) efficiently than the government—other seemingly comparable dimensions would include the range of permissible contracts, coordination of trades, production of information, responsiveness of supply and demand to money illusion, and so on. However, it is possible to produce Keynesian policy conclusions in models that incorporate rational expectations, but which contain some other departures from sensible behavior, for example, arbitrarily fixed nominal wages, or money illusion either in supply and demand functions or in the form of private labor contracts. Thus the nature of the formation of expectations seems to be an important issue within the general context of the efficiency of private arrangements relative to governmental actions, but it is this general concept of relative efficiency that seems to be crucial in evaluations of policy activism.

Monetary control is one area in which a strong governmental role is generally accepted. I abstract here from important issues that involve the transactions benefits of a generalized medium of exchange and the resource costs that can be saved by using a fiat standard for money, rather than a commodity standard. Even with this abstraction, a rationale for government involvement in monetary control can be constructed on business cycle grounds—in particular, from a consideration of the Phillips curve, which I view here as a representation of the responsiveness of economic activity to unanticipated movements in money and the absolute price level. The potential for confusion between absolute and relative price changes in a monetary economy, which is a possible basis for the Phillips curve, seems to justify some public control over the quantity of nominal money.

A major empirical finding is the central role that monetary shocks have, in fact, played in business cycles. The greater year-to-year stability of the underlying monetary mechanism in the United States since World War II—which involves a substantial increase in government regulation—has led to a smaller amplitude of business fluctuations in comparison to those of either the interwar

period or the pre-World War I era. However, it also seems that the important change in monetary structure has been a reduction in the short-run variance of money, rather than a move toward an activist feedback policy by the monetary authority. Specifically, the tendency to increase the money stock at a higher rate in response to a recession—which is a pattern that appears to originate in the Full Employment Act period following World War II—does not seem to have contributed to enhanced economic performance.

The effects of the shift to greater year-to-year stability in money indicates the potential real effects of changes in the underlying monetary institutions. The evaluation of this particular change (i.e., a comparison of the gold standard money process before World War I with today's fiat standard managed by the Federal Reserve) involves a complicated tradeoff. In my view the benefits of greater year-to-year stability in today's output and employment have been bought with two major costs. The first of these concerns the monetary education during the early years of the Federal Reserve, a process that is doubtless still continuing. The monetary errors of the interwar period—which seem to be much more extreme than those that would have arisen under the pre-World War I regime—can be credited with much of the costs of the Great Depression, as well as with those of the 1937–38 contraction. Although the gold standard was not "ideal," it did require less knowledge by the government or anyone else about how the aggregate economy worked. The second cost of the monetary change is the chronic inflation of the present environment.

Further economic benefits could be attained by moving to a monetary institution that first, and most importantly, delivered even greater year-to-year stability in the money supply, for example, by constraining the monetary authority to achieve an approximately constant growth rate for M_1, and second, that attained an average monetary growth rate below the roughly 6–7 percent annual rate for M_1 that is implied by the present structure. The precise design of the new governmental institution and, more importantly, the political-economic process that leads to changes in this or other institutions are unclear to me. However, it might be worth remarking, from the standpoint of the adjustment costs implied by a shift in monetary structure, that an analysis of expectational changes associated with a structural shift cannot usefully be separated from an analysis of the changes in the underlying variables that led to the shift in structure. In particular, there is no reason to believe that expectations about money growth and inflation are either more or less flexible than the underlying structure that generates the actual values of money growth and inflation. Therefore, one would not predict that a shift to a new monetary environment—such as those that occurred in the past concerning the monetary role of gold or the Federal Reserve—would involve an adjustment period in which expectations lagged behind the changes in "reality." A shift in the United States to an institution that delivered a lower average growth rate of money would not imply a transition period of unusually high unemployment.

The interpretation of the Great Depression is a key matter dividing policy activitists from nonactivists. The activist view is that the Great Depression was a symptom of an inherently unstable private economy that experienced large gyrations in output and which tolerated prolonged periods of high unemployment. Governmental activism in the middle and late 1930's—notably, the high levels of public expenditures—are thought to have been helpful in restoring some measure of economic prosperity.

The alternative view is that the Great Depression was in large part a product of governmental mistakes; specifically, the inept monetary policy of the Federal Reserve. Further, the governmental interventions associated with the New Deal, including the volume of public expenditures and direct price regulations, retarded the recovery of the economy, which was nevertheless rapid after 1933. The sharp increase in reserve requirements was primarily responsible for the 1937–38 recession.

I do not presently see a fully satisfactory "equilibrium" or "disequilibrium" story of the Great Depression. With respect to one type of equilibrium model that has been

proposed, I do not know of an empirical documentation of a major downward shift in factor supply during the early 1930's in response to a cut in prices relative to perceived prices. Actually, the depression experience does not stand out in this respect, since this type of factor-supply story has not been satisfactorily documented empirically even for the more mild fluctuations of the post-World War II period. On the other hand, the disequilibrium type of model, which relies on a nontheory of price rigidities, does not seem to have more impressive empirical support.

From a reduced-form perspective that relates business fluctuations to prior monetary disturbances, the contraction from 1930 to 1933 seems to be well in line with other experiences. The unprecedented monetary collapse over this period accords quantitatively with the drastic decline in economic activity. The magnitude of the monetary contraction can be appreciated by noting that the reduction in (annual average) M_1, 1929–33, averaged 7.3 percent per year. The only other four-year periods since the Civil War that show a decline in the money stock are much milder: 1875–79, 0.8 percent per year for M_2; and 1892–96: 0.4 percent per year for M_2. (A decline of 10.2 percent in M_1 occurred 1920–21, but the 1920–24 period shows no net change in the money stock.)

A somewhat greater puzzle is posed by the recovery periods 1933–36 and 1938–41. These periods may have exhibited a slower recovery rate than would have been anticipated, although the average annual growth rates of real *GNP* from 1933 to 1936 of 10.3 percent and from 1938 to 1941 of 10.4 percent (which would be reduced somewhat if the final date were 1940) are exceptional for peacetime periods. In fact, these growth rates are about the same as that (10.5 percent per year) that appears in the reported statistics for the World War II expansion 1941–44. It is possible that the recovery periods from 1933 to 1941 involve the retarding influence of massive governmental interventions into the price-setting process. However, this explanation must be regarded as highly tentative, especially since governmental interventions in the form of price controls during the Korean War and from 1971 to 1973 did not seem to have such important output effects.

In any case I regard empirical testing of hypotheses derived from alternative theories of business cycles as a matter of continuing priority. Examples of testable hypotheses that are implied by some "equilibrium" macro models with incomplete information are: 1) only the unanticipated part of movements in money affects "real" variables like output and real interest rates; 2) the anticipated part of current money movements affects the price level contemporaneously on a one-to-one basis, while the unanticipated part has a less than one-to-one contemporaneous effect on prices, a *positive* effect on anticipated real rates of return, and an ambiguous effect on nominal interest rates; 3) an increase in the variance of money would reduce the sensitivity of output to a given size money shock and raise the dispersion of relative prices; and 4) changes in consumption or in federal tax rates are, as a first-order approximation, unpredictable from lagged data. The equilibrium-type models are consistent with persisting effects of monetary and real shocks, although this aspect of the theories is less developed. The approach is also consistent with real effects of government spending, which would depend especially on the substitutability between public and private expenditures in utility and production functions, and with real, relative price-type effects of changes in taxes, unemployment compensation, and the like. Aggregate demand effects of shifts between taxes and debt issue would arise, at most, when these shifts were unanticipated.

Testable hypotheses from simple Keynesian models would seem to include: 1) increases in money imply increases in output and *decreases* in nominal and real interest rates, all of which persist over a substantial period. Price responses depend on the initial state of excess demand (whatever that means), but generally the rate of change of prices is raised above what it otherwise would have been. However, the short-run effect on prices is weak. The role of price and money expectations is not stressed in simple models of this type. 2) Increases in the government

deficit, produced by higher expenditures or reduced taxes with the money stock held fixed, imply increases in output—which are likely to be multiplicative—and increases in interest rates. These effects persist over a substantial period. Prices rise at a faster rate than otherwise, but the initial price response is weak. The nature of government expenditures and expectations about future taxes, prices, etc., are not stressed in simple models of this type. 3) Real wages move countercyclically in models that assume only wage stickiness. In models that also assume some price stickiness, the cyclical behavior of real wages is indeterminate.

I will not attempt at this time to present a detailed appraisal of the state of current empirical evidence—however, I think that the major doubts about Keynesian, non-market-clearing-type analysis that are prevalent today are primarily a reflection of perceived empirical inadequacies of the theory, especially in an inflationary environment. I expect that the final verdict on the usefulness of Keynesian economics will also come primarily from empirical analysis.

Evaluating the Non-Market-Clearing Approach

By PETER HOWITT*

This paper is concerned with evaluating the non-market-clearing (*NMC*) approach of Robert Barro and Herschel Grossman and others from a purely positive point of view. That is, it deals with the broad question of the extent to which the approach provides a theoretically satisfactory explanation of certain stylized facts characterizing the dynamic behavior of aggregate output and the price level. It does not deal with the important and difficult normative questions involving stabilization policy that are often associated with the approach.

From this viewpoint the main strength of the *NMC* approach is its compatibility with the evidence that 1) fluctuations in aggregate output are closely (positively) correlated with fluctuations in aggregate demand, 2) output appears to respond with a much shorter lag than does the price level to changes in aggregate demand, and 3) changes in output are serially correlated from quarter to quarter.

The main weakness of the approach is its failure to provide any satisfactory account of how markets are organized. For example, it offers no explanation of how prices are formed, beyond the crude hypothesis that they move in the direction of excess demands, despite the fact that the assumption that prices fail to respond quickly enough to clear markets lies at the heart of the approach. Nor does it explain why agents should be constrained to trade at these prices, even though these constraints are what ultimately produce the multiplier process of the approach. This inattention to the details of market organization also appears to be responsible for the curious "supply multiplier," according to which an increase in aggregate demand, from an initial position of generalized excess demand or even of full employment equilibrium, causes a decrease in output—a prediction that threatens to undermine the compatibility of the approach with the positive correlation between aggregate demand and output unless some reason can be found why excess demand should be less common than excess supply.

This shortcoming does not imply that the *NMC* approach is not useful for many purposes, nor that its predictions are inconsistent with the evidence (except for the predictions of the supply multipliers). But to be consistent with the evidence is not to explain it. What the approach lacks is a satisfactory theoretical underpinning that would at least make it consistent with the same notions of rational self-interest that underlie the rest of economic theory.

This leaves us with the question of whether a satisfactory underpinning can be provided to the approach. In other words, can the approach be revised or replaced in such a way that the resulting theory contains a more satisfactory account of market organization, and explains the above mentioned stylized facts in a way that closely resembles the *NMC* approach.

This question cannot now be answered with a great deal of confidence because no one has yet developed a satisfactory theory of market organization. However I think that an affirmative answer is likely, and that the key to developing the answer lies in recognizing that different markets are organized in different ways. In particular some markets, such as those for many labor services, personal credit, and heavy capital goods, are organized on a highly personal basis with individually negotiated contracts, whereas other markets, such as those for widely traded financial assets and for most consumer durables, are organized on a less personal basis by trading specialists like retailers, wholesalers, jobbers, brokers, and stock market specialists. The rest of this paper attempts to shed some light on the question of providing a satisfactory theoretical underpinning for the *NMC* approach by investigating how a market organized by such specialist

*University of Western Ontario. I am indebted to David Laidler for helpful conversations on the topic of this paper, to Robert Solow for his critical comments, and to the Humanities and Social Sciences Research Council of Canada for financial support

traders would behave following a reduction in aggregate demand.

In such a market all trading activity is typically arranged by inventory-holding traders who quote prices at which they stand willing to sell to demanders at dates and in quantities that may be chosen by the buyer. Some traders also quote prices at which they likewise stand willing to buy from suppliers at the suppliers' discretion, as in the case of stock market specialists, although this arrangement is less common. These traders may be separate middlemen who neither produce nor directly consume the good being traded; or they may be identical with the producers of the good, as in the case of the sales departments of large manufacturing companies that produce to stock. In order to back up their willingness to trade, specialist traders allow their inventories to act as buffer stocks so as to absorb unanticipated fluctuations in excess demand, except in cases where the fluctuations are large enough to exhaust either the trader's inventories or his storage capacity. Thus under normal circumstances they allow other traders (at least buyers) to make their notional plans effective, even when full general equilibrium prices have not yet been established.

The existence of such specialist traders can be rationalized by supposing that there are costs of transacting, which the specialists reduce by being available for trading at known locations and at announced prices, and by their willingness to refrain from quantity rationing. Because there is a large set up component to the cost of transacting—a fixed cost paid at each transaction date regardless of the amount traded—agents will choose to trade at discrete dates rather than continuously in time. This means that some device is needed to coordinate their timing decisions. Specialists partly fulfill this role by standing open continuously for business, thereby permitting others to make independent timing decisions. Thus the specialist's inventories absorb not only the unanticipated fluctuations in excess demand that will eventually call for a price change, but also the transitory fluctuations due to differences in arrival times of buyers and sellers (see Robert Clower and the author).

How such a market would react to a decline in aggregate demand obviously depends upon what kind of pricing policy is pursued by the specialists. The theory of optimal price-setting behavior is only beginning to be worked out. Meanwhile an intuitive and simple working hypothesis is to suppose that prices are always set at their expected market-clearing values. That is, each specialist's selling price is one such that over any interval of time the specialist expects to sell the quantity that he would prefer to sell if in fact he could choose the volume of his sales at that price.

Suppose that prices are set at their expected market-clearing values. When aggregate demand falls, if this is anticipated in all markets and if each specialist is able to anticipate its effects on all other prices, then all prices will fall immediately so as to offset the expected effects of the shock. But generally it will not be fully anticipated. At first it is more likely to be interpreted at least partly as a temporary change in demand. If so, then the price level at first will not respond to the shock, for the same reason that in any speculative market the price is unlikely to respond to a temporary shock in demand or supply. Instead, the specialists, like speculators, will allow their inventories to accumulate to absorb the shock. At first the rate of output may also remain unchanged, but as the level of inventories increases the rate of output is likely to decrease. How this happens is most easily seen in markets where producers are their own specialist traders. As the level of inventories increases, the relative shadow price attached by a firm to its inventories is likely to decline, thereby inducing a lower rate of production. Thus right away there is a strong resemblance to the *NMC* approach in the sense that output will move in sympathy with aggregate demand, except when there is perfect price flexibility (which occurs when the shock is fully anticipated). It is unlike most versions of the *NMC* approach, however, in that the connection between aggregate demand and output involves the same sort of inventory change that is a common feature of textbook accounts of the Keynesian cross diagram.

An important feature of the *NMC*

approach is the multiplier process, according to which the ultimate decline in output may exceed the initial decline in aggregate demand, provided that the deviation-counteracting feedback effects of price reductions on aggregate demand are not large enough to offset the deviation-amplifying effects of quantity rationing. This too is likely to be consistent with an aggregate model in which output markets are assumed to be organized by specialists. Indeed the multiplier process is one that does not depend, as is often supposed, upon the failure of markets to clear. All it requires is that the reduction in output and employment be associated with a reduction in the typical household's expected lifetime wealth, causing a secondary reduction in aggregate demand that is not fully anticipated by all price setters. In the *NMC* approach this shift occurs as a result of the quantity rationing of sellers. But even in a market-clearing approach in which the connection between aggregate demand and output arises from the inability of individuals to distinguish a change in aggregate demand from a change in relative demand, the same perceived change in relative demand that induces quantity reductions will also induce a reduction in the typical household's expected lifetime wealth. The same is true of models in which the labor market is organized around long-term contracts. The worker who is laid off by a firm that partly interprets the decline in aggregate demand as a decline in relative demand does not himself have to suffer from any such confusion to reduce his estimate of his lifetime wealth.

An important sense in which the present approach differs from the *NMC* approach is that it does not rely in any essential way upon the notion of quantity rationing. In labor markets employment will respond to a balanced reduction in aggregate demand if and only if it is at least partly interpreted in some markets as a reduction in relative demands. Whether separations in the labor market take the form of quits or layoffs does not matter for any aggregate predictions.

On the other hand, in output markets the price-setting behavior of the specialists appears to be consistent with the market-clearing (*MC*) approach of Robert Lucas and others. The situation of a specialist and a customer who has just appeared in his store may be treated as a "Phelpsian island" informationally isolated from other such islands. According to the *MC* approach the price on such an island should equal the temporary equilibrium price that clears the market, given each side's expectations of the market-clearing price on the next island. It seems reasonable to suppose that the specialist's supply curve in such a temporary equilibrium situation would be horizontal, or nearly so, at a price equal to the expected equilibrium price, for this represents the expected opportunity cost of selling on this island rather than the next. If this were true then the *MC* approach would result in exactly the same behavior as the present revised version of the *NMC* approach, and the inventory accumulation that appears as "involuntary" to the textbook Keynesian accounts of the *NMC* approach can be regarded as "voluntary" speculation in the present approach.

The point of all this is not to argue that what Keynesians regard as involuntary is really voluntary, but rather to argue that the *NMC* approach does not depend in any essential way upon a violation of the basic assumption of rational self-interest in the form of a failure to exploit perceived gains from trade. Indeed the above discussion suggests that the distinction between voluntary and involuntary behavior is not a useful one for macroeconomic theory. The misery of unemployment is as great if it is voluntary or involuntary, the behavior of the unemployed does not depend upon whether they quit or were fired, and the reaction of a firm to an accumulation of inventories does not depend upon whether the accumulation was involuntary or whether it is a result of voluntary speculation. If progress is to be made in unifying the micro and macro branches of economic theory it is vital that we be able to base macro theory upon principles of voluntary behavior. I think this is possible without substantially altering the character of orthodox macro theory. But if I am right it is probably simpler in most applications of macro theory to assume stickiness of prices and to assume that markets fail

to clear than to bother specifying the entire set of assumptions that would make this consistent with rational self-interest.

REFERENCES

Robert J. Barro and Herschel I. Grossman, *Money, Employment, and Inflation*, New York 1976.

R. W. Clower and P. W. Howitt, "The Transactions Theory of the Demand for Money: A Reconsideration," *J. Polit. Econ.*, June 1978, *86*, 449–66.

R. E. Lucas, "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

Why Does Aggregate Employment Fluctuate?

By HERSCHEL I. GROSSMAN*

Business cycles appear to reflect predominantly the effects of changes in aggregate demand for output on real variables such as aggregate employment. The measurement of the magnitude and timing of these effects is not a trivial problem. However, recent work, for example, by Robert Hall and by Robert King, supports the impression that the response of aggregate employment to aggregate demand is prompt and large, apparently peaking within a few months of the disturbance at a level equalling or exceeding the value of the disturbance, but also temporary, giving way over time to adjustments in wages and prices.

These observations suggest that cyclical fluctuations in aggregate employment are not symptoms of the familiar nonneutralities that arise in analyzing the effects of changes in the stock of money or its velocity of circulation in a Walrasian general equilibrium context. Rather, the characteristics of the relation between aggregate demand and aggregate employment suggest that the actual economy differs in significant respects from the imaginary Walrasian economy, in which exchange takes place only under market-clearing conditions that emerge from a tâtonnement process conducted with all agents in possession of complete information about wages and prices.

This conclusion that the Walrasian paradigm does not do justice to the real world does not seem controversial. However, research in macroeconomics has witnessed persistent disagreement about the usefulness of specific non-Walrasian paradigms for modelling the determination of aggregate employment. This disagreement reflects basic differences in perceptions of the essential characteristics of the actual economy that are responsible for the non-Walrasian behavior of aggregate employment.

The discussion that follows evaluates the current state of this controversy, focusing on recent disenchantment with the popular non-market-clearing approach and, more positively, on the prospect that a resolution of this paradigm conflict is presently at hand. As explained below, this impending settlement involves recognition that incomplete information is the critical factor in the generation of non-Walrasian fluctuations in employment, together with explicit allowance for the implications of implicit contractual arrangements for efficient shifting and pooling of risk in labor and product markets.

The present discussion is not concerned with the causes of fluctuations in aggregate demand or, in particular, with the relative importance of variations in the stock of money and its velocity of circulation. Rather, the focus is only on why such fluctuations in aggregate demand cause observed non-Walrasian fluctuations in aggregate employment, taking this causal link to be a matter of fact.

I. Non-Market-Clearing Paradigm

In recent years theoretical models of macro-economic relations have been judged not only by their consistency with observed phenomena, but also by their use of convincing choice-theoretic rationalizations for the underlying behavior of individuals. Over the past decade or more, these criteria have motivated substantial research aimed at reworking the theory of macro-economic relations by using the same foundations and the same principles that have traditionally served in theorizing about micro-economic relations. Included in this body of research is the work that develops the non-market-clearing paradigm, according to which the causal relation between aggregate demand and employment results from the failure of wages and prices to

*Brown University. The National Science Foundation has supported this research. Robert Gordon, Robert King, and William Poole have given me helpful comments.

adjust to equate quantities demanded and supplied in labor and product markets.

The primary attraction of this paradigm has been that it explicitly takes into account the allegedly obvious fact that actual markets for most labor services and many products chronically fail to clear. This purported observation is based on such phenomena as the prevalence of layoffs and other apparent symptoms of non-wage rationing of employment and the lack of a pronounced cyclical pattern in measured real wage rates. An essential aspect of the non-market-clearing paradigm is that a contraction in employment resulting from a reduction in aggregate demand involves a situation in which perceived gains from trade are foregone because buyers and sellers are limited to transacting at a wage-price vector that does not equate quantities supplied and demanded.

As many writers have stressed, including Robert Barro and myself in the introduction to *Money, Employment and Inflation*, the theoretical development in the existing literature provides no convincing rationale for such a persistent restriction on transactions. For example, in our book the determination of wages and prices is based on *ad hoc* gradual adjustment processes, and the choice-theoretic analysis is concerned mainly with the implications of such essentially arbitrarily specified wage-price vectors for the determination of employment. Some other models rationalize gradual wage and price adjustment on the basis of adjustment costs, which is logically adequate, but convincing stories about the precise nature of these costs do not seem to exist.

The problem of explaining the failure of markets to clear does not arise in situations in which legal restrictions prevent wage and price adjustments. In fact, one useful product of the non-market-clearing paradigm has been the model of the supply multiplier, which Barro and I developed to analyze suppressed inflation. Studies of centrally planned economies, such as the work of David Howard, have implemented this model empirically with some success. However, legal restrictions on wage and price adjustments surely do not play an important role in the determination of aggregate employment in contemporary Western economies. Consequently, the lack of a model of the market process that provides choice-theoretic underpinnings for the failure of markets to clear presents a relevant and basic problem for the non-market-clearing paradigm.

A few years ago, a continuing effort to develop such a model was high on my own research agenda. At that time, the observation that markets chronically fail to clear appeared to me so obviously correct that an effort to fill this missing link in the microeconomic foundations of the non-market-clearing paradigm seemed to be the only defensible research strategy. More recently, however, a theoretical innovation based on the idea that labor market transactions involve largely implicit contractual arrangements for shifting risk from workers to employers has led to models that rationalize the observed stickiness of measured real wage rates and explain the alleged symptoms of non-wage rationing of employment without invoking the failure of markets to clear. Before discussing these models of implicit contracts, it is useful to consider the paradigm of incomplete information, which is the principal alternative to the non-market-clearing paradigm for analyzing the causal relation between aggregate demand and employment.

II. Paradigm of Incomplete Information

In the existing literature, the development of the incomplete-information paradigm within the framework of competitive markets has used three different, but mutually consistent, stories. One story, which Milton Friedman tells, is that the worker typically does not have complete information about the prices of the items that he consumes, and, hence, he tends to overestimate the extent to which a change in the nominal value of his product, as signalled by his nominal wage rate, involves a change in his terms of trade between leisure and consumption. A second story, which Dale Mortensen develops, is that the worker typically does not have complete information about wage rates at alternative places of

employment, and, hence, changes in wage rates cause him to misjudge the probable returns from continued job search. This story is likely to apply mainly to workers who are not currently employed. A third story, which Robert Lucas and Leonard Rapping develop and Lucas and Barro have subsequently used, is that the worker typically does not know whether a change in his wage rate is permanent or transitory, and, hence, tends to overestimate the extent to which changes resulting from nominal disturbances, which are assumed to be permanent, involve changes in the terms of trade between current and future leisure, as measured by the real rate of interest. In all of these stories, changes in aggregate demand affect aggregate employment because individuals, especially labor suppliers, do not have sufficient information to distinguish clearly the price and wage signals transmitted by such disturbances from the price and wage signals associated with shifts in the pattern of demand. Buyers and sellers do not forego perceived gains from trade, as in the non-market-clearing paradigm, but incomplete information causes them to perceive potential gains from trade incorrectly.

Models of conjectural equilibrium, analyzed by Frank Hahn and Takashi Negishi, represent another line of development of the incomplete-information paradigm, within a framework of monopolistic markets. In these models, price-setting transactors act to maximize expected utility based on conjectures about demand curves rather than knowledge of true demand curves.

The incomplete-information paradigm is attractive primarily because, developed in terms of any of the above stories, it seems to offer a fully choice-theoretic explanation for the relation between aggregate demand and aggregate employment. These stories, however, presumably require the underlying rationalization that for individuals obtaining complete relevant information about, for example, prices or monetary aggregates is on average too costly to be worthwhile. Accordingly, acceptance of the incomplete-information paradigm, together with rejection of the non-market-clearing paradigm, implies a belief that significant costs of obtaining relevant information are plausible, whereas significant costs of wage and price adjustment are not. In the current state of knowledge, these impressions about costs, however sound, admittedly derive only from casual evidence.

A frequent objection to the incomplete-information paradigm is that it cannot readily account for observed persistence in the effect of shifts in aggregate demand on aggregate employment. However, as Lucas and Thomas Sargent stress, an absence of serial correlation in misperceptions of potential gains from trade does not preclude serial correlation in the effects of such misperceptions. Moreover, the alternative non-market-clearing paradigm does not seem to have any basic advantage with respect to explaining persistence. Specifically, assuming that wages and prices adjust gradually so that excess supply is persistent seems no less heroic than assuming that information disperses gradually or, as Lucas and Sargent suggest, that demands for labor services or physical capital adjust gradually. Finally, the recent studies by Hall and King indicate that the amount of persistence in unemployment and employment is actually much less than a casual observer might think.

Another objection to the stories about incomplete information, which at one time made the incomplete-information paradigm seem less attractive to me than the non-market-clearing paradigm, is that prominent qualitative aspects of the relation that these stories predict between aggregate demand and aggregate employment appear to be empirically unacceptable. Specifically, as originally formalized within a framework of spot markets, these stories make no allowance for symptoms of non-wage rationing of employment, such as layoffs, and they also predict that cyclical variations in employment involve countercyclical variation in quit rates and real wage rates. Note that agents in the monopoly models of Hahn and Negishi face quantity constraints only in the sense that for a monopolist profit-maximizing price exceeds marginal cost. In addition, the two stories that involve misperceptions of the relation between current nominal wage rates and either current prices or future wage rates are subject to a theoretical reservation because, in order for

the predicted effect of aggregate demand on aggregate employment to be positive, they require sufficient restrictions on worker utility functions to make substitution effects strongly dominant over income effects. These observations suggest that the stories about incomplete information, as originally formulated, achieve rigor only at the sacrifice of essential realism and/or theoretical generality.

III. Implicit Contractual Arrangements to Mitigate Risk

As mentioned above, recent development in the theory of risk shifting in labor markets, and related extensions to product markets, offer a possible way to resolve this dilemma. According to the risk-shifting theory, relations between firm and workers implicitly involve two transactions. First, firms purchase from workers labor services for use in the production process and, second, firms sell to workers insurance against undesirable income fluctuations. As a result of these insurance arrangements, a worker's nominal wage income equals either the value of his marginal product minus an implicit insurance premium or the value of his marginal product plus an implicit insurance indemnity, depending on whether the perceived real value of his marginal product is high or low. In line with the above stories about incomplete information, the perceived real value of marginal product refers to the current nominal marginal product relative to perceptions of either current prices or future marginal products. It is important to note that this formulation presumes that firms and workers share the same perceptions, but explicitly allows for the possibility that these perceptions can be incorrect.

In some recent papers, I have explored the usefulness of this particular contractual view of labor markets for understanding the relation between aggregate demand and employment. My 1978a,b papers develop explanations for the phenomena associated with layoffs without reference to a failure of labor markets to clear and a loss of perceived gains from trade. In these papers, observed unresponsiveness of measured wage rates to demand disturbances is a result of efficient risk-shifting arrangements. Differences in the stability of earnings between more and less senior workers result from differences in productivity and either differences in reliability, which concerns willingness to work for less than the value of product when the perceived real value of product is high, or the availability of other subsidized income sources such as unemployment insurance.

Most importantly in the present context, stickiness of wage rates in this analysis is not a causal factor connecting aggregate demand and employment. Productive efficiency conditions, which, as Barro has stressed, are a consequence of competition in the markets for labor contracts, imply that a worker is employed in a particular state of nature if the utility associated with being employed and receiving the perceived real value of his marginal contribution to total product equals or exceeds the utility associated with not being employed. This analysis implies that, although a laid-off worker would want to work if offered either the wage rate he received when he was employed or the wage rate inclusive of insurance indemnity received currently by more senior workers who are employed, he would typically not want to work at the wage rate that would currently clear a spot market for labor services, given perceptions of current prices and future marginal products. Thus, the use of layoffs to effect employment separations does not imply that the amount of employment is suboptimal relative to current perceptions. This theory allows aggregate employment to differ from the hypothetical Walrasian outcome, but such differences—as modelled, for example, in a recent paper by Costas Azariadis—reflect misperceptions of either current prices or future marginal products, as in the above stories about incomplete information.

My 1979a paper extends the idea of implicit contractual arrangements to the pooling of risk in product markets. In this more complete model, because the bulk of transactions take place at contractually predetermined wages and prices, large changes in output and employment can occur without large changes in average wage and price levels. Thus, even though changes in spot

product prices are important in this analysis, the hypothesis that risk-mitigating arrangements are prevalent implies that the data should exhibit at most a weak correlation between aggregate employment and average real wage rates. Finally, my 1979b paper shows that the existence of risk-shifting arrangements in labor markets strengthens the substitution effects that influence the choice of the efficient level of employment, implying that only weak restrictions on worker utility functions are necessary for changes in current nominal marginal products relative to perceived prices or expected future marginal products to have a strongly positive effect in employment.

IV. Conclusions

These recent results indicate how a view of market transactions including implicit contractual arrangements for mitigating risk can reconcile the alleged facts—specifically, apparent non-wage rationing of employment and stickiness of real wage rates—with the notion that contractions in employment resulting from reductions in aggregate demand involve incorrect perception of gains from trade, as implied by the incomplete-information paradigm. Thus, it now seems that the hypothesis that non-Walrasian fluctuations in employment result basically from the limited ability of economic agents to distinguish aggregate disturbances from relative disturbances, when incorporated into a framework of contractual rather than spot markets, can provide a basis for analyzing the relation between aggregate demand and aggregate employment that is both fully choice theoretic as well as consistent with observed characteristics of employment fluctuations.

What verdict does this discussion suggest regarding the usefulness of the non-market-clearing paradigm? On the one hand, the results summarized above suggest that the alleged facts do not provide unambiguous support for the view that the essential problem is failure to realize perceived gains from trade, as implied by the non-market-clearing paradigm. On the other hand, existing research has not formulated testable hypotheses and used systematic empirical evidence to reject the non-market-clearing paradigm. Moreover, as mentioned above, this paradigm has clear relevance for analyzing specific situations in which legal restrictions prevent wage or price adjustments. More generally, this paradigm provides an account of employment fluctuations that is simple and convenient for expository purposes, although the above discussion suggests caution in drawing normative implications from models that do not emphasize the factor of incomplete information. Nevertheless, the main implication of the present discussion is that, as a basis for a general theory of the causal relation between aggregate demand and aggregate employment, the non-market-clearing paradigm is less attractive than it once seemed, especially in comparison to the alternative paradigm of incomplete information extended to take account of implicit contractual arrangements for mitigating risk.

REFERENCES

C. Azariadis, "Escalator Clauses and the Allocation of Cyclical Risks," *J. Econ. Theory*, June 1978, *18*, 119–55.

Robert Barro, "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

———, "Long-Term Contracting, Sticky Prices, and Monetary Policy," *J. Monet Econ.*, July 1977, *3*, 305–16.

——— **and Herschel I. Grossman**, *Money, Employment, and Inflation*, New York 1976.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

H. I. Grossman, (1978a) "Risk Shifting, Layoffs, and Seniority," *J. Monet. Econ.*, Nov. 1978, *4*, 661–86.

———, (1978b) "Risk Shifting, the Dole, and Layoffs," unpublished paper, Brown Univ., Aug. 1978.

———, (1979a) "Employment Fluctuations and the Mitigation of Risk," *Econ. Inquiry*, Oct. 1979, forthcoming.

———, (1979b) "Incomplete Information,

Risk Shifting, and Employment Fluctuations," unpublished paper, Brown Univ. 1979.

R. E. **Hall**, "Expectation Errors, Unemployment, and Wage Inflation," unpublished paper, Center Advanced Stud. Behav. Sci., Stanford Univ., Sept. 1977.

F. H. **Hahn**, "On Non-Walrasian Equilibria," *Rev. Econ. Stud.*, Feb. 1978, *45*, 1–17.

D. H. **Howard**, "The Disequilibrium Model in a Controlled Economy: An Empirical Test of the Barro-Grossman Model," *Amer. Econ. Rev.*, Dec. 1976, *66*, 871–79.

R. **King**, "Dynamics of Unemployment and Employment: A Reevaluation of the U.S. Time Series, 1950–1976," unpublished paper, Brown Univ. 1978.

R. E. **Lucas, Jr.**, "Understanding Business Cycles," in Karl Brunner and Allan H. Meltzer, eds., *Stabilization of the Domestic and International Economy*, New York 1977.

——— **and L. A. Rapping**, "Real Wages, Employment, and Inflation," in Edmund S. Phelps, et. al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

——— **and T. J. Sargent**, "After Keynesian Economics," unpublished paper, Univ. Chicago and Univ. Minnesota, July 1978.

D. T. **Mortensen**, "A Theory of Wage and Employment Dynamics," in Edmund S. Phelps, et. al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

T. **Negishi**, "Existence of an Under-Employment Equilibrium," in Gerhard Schwödiauer, ed., *Equilibrium and Disequilibrium in Economic Theory*, Boston 1978.

The Productivity of Foreign Resource Inflow to the Soviet Economy

By PADMA DESAI*

The recent growth in Soviet borrowing from the West, resulting in the rise of net Soviet debt to $16 billion by 1977, has raised the analytical question of considerable economic and political significance: how much are such credits worth to the Soviet Union? Several alternative approaches to answering this question are possible. Broadly speaking, they may be grouped into two categories: partial equilibrium and general equilibrium.

In the partial equilibrium analyses, we may include (i) the econometric investigation of whether foreign capital goods have greater productivity in Soviet industry and branches than domestic capital goods, the relevance of this to the problem at hand being established by the argument that foreign credits may be jointly supplied with the capital goods embodying advanced technology; and (ii) the noneconometric examinations of the role that foreign credits and technology can or do play in specific sectors (for example, automobiles and oil). On the other hand, a partial equilibrium approach is not a satisfactory way to approach the question at hand. The productivity of foreign credits can be defined meaningfully only in the context of the use of resources generally, requiring a general equilibrium approach.

In turn, a general equilibrium approach may be based on the incorporation of foreign credits into one of the "large" models, either of the computable planning type or of the econometric variety as represented by the *SOVMOD* exercise of Donald Green and Christopher Higgins, the productivity of foreign credits being estimated by the required variation in the levels of foreign borrowing therein. However, these large models tend to be rather cumbersome and the total effects of parametric variations are therefore extremely difficult to disentangle. "Small" models, by contrast, often have the advantage of both elegance and ease of interpretation, while not burying the essence of the economic system within a complex structure.

In this paper, I have chosen the small, "simple decision" model approach and elaborated a general-equilibrium model of the Solow-Swan variety, using a *CES* production function. I have estimated the functional relationships postulated and calculated the productivity of foreign resource inflow (of which foreign credits are a component) within it. In Section I, the model is set out, its structure is justified, and the analytical methods to be used to calculate the productivity of foreign resource inflow in it are spelled out. In Section II, I present the estimates of the model and the productivity of foreign resource inflow that emerges from them. Section III offers some concluding observations, in light of the estimates.

I. The Model and its Rationale

A. *Underlying Rationale of the Model*

The model developed here is of the Solow-Swan variety. One of its key characteristics, as is well known, is that it is a "flow" model, where savings define investment and hence

*Research associate, Russian Research Center, Harvard University, and visiting professor of economics, Boston University. Thanks are due to the National Science Foundation Grant No. 77-07254 for partial financial support of the research underlying this paper. The excellent research assistance provided by Ricardo Martin was not merely competent but creative and has contributed greatly to the writing of this paper. A longer version is available on request.

the growth of the economy given the marginal capital-output ratio, in contrast to a typical "structural" model of the Feldman-Mahalanobis variety where the current investment allocation pattern defines the feasible savings in the future and hence the growth of this economy.

This representation of the Soviet economy is not unusual. In a simpler and closed version, it is essentially the underlying model of Abram Bergson's (1971, 1973) influential analysis of Soviet growth prospects. But the use of such a model for the problem at hand does imply that the constraint on the *growth* of the Soviet economy is provided by savings and that the productivity of foreign resource inflow is determined by its direct addition to domestic investment, the latter being always equal to the sum of domestic savings and foreign resource inflow.

This view therefore ignores the possible role of foreign resource inflow in breaking a "foreign exchange" or "transformation" bottleneck to Soviet growth. If foreign exchange is considered to be the constraint on Soviet growth, then clearly the estimate of the productivity of foreign resource inflow in this paper should be considered to be a "lower bound" estimate.

B. *The Model*

The model consists of the following equations, written in their estimating form:

(1)
$$Y_t = Ae^{\lambda t}[\alpha K_t^{*-\rho} + (1-\alpha)L_t^{-\rho}]^{-1/\rho} \cdot e^{u_t}$$

where $A > 0, 0 \leq \alpha \leq 1$ and $\rho > -1$. This is the aggregate *CES* production function for the Soviet economy, with Hicks-neutral technical change (at rate λ); Y_t is *GNP*, L_t is total employment in man-hours, K_t^* is the average capital stock (during year t), t is time in years, and u_t is the error term.

$$(2) \qquad S_t = I_t - F_t$$

This is the basic savings-investment identity where S_t is domestic savings, I_t is gross fixed investments, and F_t is foreign saving, measured as the excess of Soviet imports over exports.

$$(3) \qquad S_t = B + sY_t + \epsilon_t = S_o + s(Y_t - Y_o) + \epsilon_t$$

where $0 \leq s \leq 1$. This is the Keynesian-type savings function for the Soviet economy and reflects, of course, the planners' decision on the rate of saving.

$$(4) \qquad K_{t+1} = K_t + I_t - D_t$$

$$(5) \qquad D_t = \delta K_t^* + \mu_t$$

where $0 \leq \delta \leq 1$

$$(6) \qquad K_t^* = (K_t + K_{t+1})/2$$

Equations (4)–(6) relate to the relationship between investment and capital accumulation; K_t is the capital stock at the beginning of the year t and K_{t+1}, that at the end of it; K_t^* is then defined by equation (6) as the simple average of the two, entering equation (1) as the average, effective capital stock; D_t refers to depreciation/retirement of capital, expressed as a proportion of the average capital stock in equation (5). Equation (4) will determine D_t, given the observed values for $(K_{t+1} - K_t)$ and I_t.[1]

$$(7) \qquad L_t = L_o \cdot e^{\lambda_L t}$$

Finally, the growth of employment is exogenously determined at rate λ_L.

The system stated above has seven equations. Given F_t and K_t exogenously at time t, it will determine the seven unknowns: Y_t, S_t, D_t, K_t^*, I_t, I_{t+1}, and L_t. In addition, the system is recursive, determining the evolution of the economy from initial conditions defined by F_o and K_o as a function purely of the time path of F_t.

[1]Note that in equation (4), the addition to capital stock $(K_{t+1} - K_t)$ *should* actually be represented by the difference between activated capacity (and *not* investment) and retirements. There is also the additional problem of the lag between investment expenditures and activated capacities. Note that Soviet official sources do publish information on activated capacity (for example, *Vvod v deistvie osnovykh fondov*). However, a complex subsystem of investment and capital accumulation with the use of official estimates of activated capacity in equation (4) and also lags between investment and activated capacity did not give statistically meaningful estimates of the lag parameters nor an economically meaningful estimate of δ.

To see this clearly and to simplify the numerical solutions later, it is convenient to define a new variable: $k_t = K_t^*/L_t$, the average, effective capital-labor ratio. Then, from equations (1)–(3), we can write

$$(8)\quad B + sAe^{\lambda t}[\alpha k_t^{-\rho} + (1-\alpha)]^{-1/\rho}\cdot L_t = I_t - F_t$$

whereas from equations (4) and (5) we obtain

$$(9)\quad K_t^* = L_t\cdot k_t = (2/[2+\delta])\cdot K_t + (1/[2+\delta])\cdot I_t$$

Replacing I_t from equation (9) in equation (8), we then obtain

$$(10)\quad B + sAL_t\cdot e^{\lambda t}[\alpha k_t^{-\rho} + (1-\alpha)]^{-1/\rho} = (2+\delta)L_t\cdot k_t - 2K_t - F_t$$

Equation (10) clearly determines k_t (and hence, from equations (1) and (3), also Y_t and S_t) as a function only of K_t the capital stock at the beginning of the period t, L_t the labor supply (as given by equation (7)), and the exogenously specified F_t, the foreign resource inflow. Once S_t and F_t are known, I_t is also determined from equation (2) and so we have also the capital stock at the *end* of the period. Therefore the system can be solved recursively into the future. In fact, from equation (6), we can write

$$(11)\quad K_{t+1} = 2K_t^* - K_t = 2L_tk_t - K_t$$

Equations (10) and (11) serve as the reduced-form version of our system, with which we can perform the necessary calculations.

The calculation of productivity of foreign inflow to the Soviet Union in this framework can then be attempted in the following manner. First, for alternative "plausible" time paths of Soviet absorption of foreign resource inflow during 1978–85, we can estimate the incremental income[2] resulting from such inflows. Second, for these alternative plausible profiles of foreign resource inflows we can then estimate the "internal rate of return," represented by these incremental incomes in each of the years. We thus reduce the benefit vectors to single numbers that may be put against the cost of foreign resources to the Soviet Union.[3]

II. Estimating the Model

A. *The Production Function*

My estimates of the parameters of equation (1) with a *log* formulation and non-linear estimation procedure are as follows:

$A =$	0.7759	$\lambda =$	0.0173
	(81.2653)		(3.7555)
$\rho =$	1.1828	$\alpha =$	0.4605
	(3.7704)		(8.9826)

$R^2 = 0.9987$;

Durbin-Watson statistic (D-W) = 1.7830

(The values in parentheses are the estimated t-values of the corresponding parameters.) Note that in estimating the equation, I used data for the period 1950–1975 and measured the variables as follows:

- Y_t = Estimate of *GNP* in billion 1970 rubles by Rush Greenslade
- L_t = Total employment in billion man-hours estimated by Murray Feshbach and Stephen Rapawy
- K_t^* = Average capital stock in 1955 prices in billion rubles from official sources (see *Narkhoz*, 1956–75)
- t = time in years, t = 0 in 1950

The elasticity of substitution σ is therefore estimated as $\sigma = 0.4583$. The hypothesis that it is equal to one (i.e., that $\rho = 0$) is strongly rejected by a t-test (although of course that test can be justified only asymptotically here). So is the hypothesis of no technical change (i.e., $\lambda = 0$). Evidently, the fit is remarkably good and yields a plausible modest technical

[2]Note that incremental income streams are those resulting from alternative flows of foreign resource inflow over and above those accruing from zero resource inflow.

[3]Since we are truncating the benefit stream (i.e., incremental income stream) in 1985, the calculation of the internal rate of return will have to adjust the benefit stream by adding the value of the terminal equipment, attributable to the foreign inflow, left over at the end of 1985.

change of 1.7 percent for the economy and a low elasticity of substitution.[4]

B. *The Saving Function*

I estimated saving from equation (2): $S_t = I_t - F_t$ by using data on I, X, and M where

I_t = Estimate of gross fixed investment in billion 1970 rubles by Greenslade

X_t = Total merchandise exports f.o.b. in 1970 prices in billion rubles from official sources (see *Vneshnyaya Torgovlya*, 1967–75)

M_t = Total merchandise imports f.o.b. in 1970 prices in billion rubles from official sources

$F_t = M_t - X_t$

With these calculated saving, I estimated equation (3): $S_t = B + sY_t + \epsilon_t$ for the period 1955–75. An initial estimation showed a high degree of serial correlation in the residuals, so the equation was reestimated assuming $\epsilon_t = \rho'\epsilon(t-1) + \eta_t$.

My estimated parameters of equation (3) are as follows:

$$B = -21.5876 \quad (8.4106) \qquad s = 0.3340 \quad (38.1675)$$
$$\rho' = 0.3099 \quad (1.4576)$$
$$R^2 = 0.9939; \quad D\text{-}W = 1.7995$$

(The values in parentheses are the estimated t-values of the corresponding parameters.) The marginal propensity to save for the Soviet economy thus turns out to be 33 percent.

C. *Investment and Capital Accumulation*

Next, equation (4), $K_{t+1} = K_t + I_t - D_t$, was estimated with official data on (K_{t+1}) and on I_t, where K_t = capital stock at the beginning of the year in billion rubles; D_t = retirements of capital stock during the year in billion rubles. The data on K_t and K_{t+1} also enabled me to calculate K_t^* in equation (6).

Finally, the estimated series of D_t from equation (4), when regressed on K_t^*, gave an estimate of equation (5): $D_t = \delta K_t^* + \mu_t$. The estimated value of δ was 0.0482, with a standard error of 0.0025 and a t-value of 19.5623. (The standard error of the regression was 6.7147, the Durbin-Watson statistic was 2.2057 and the R^2 was 0.7182.) It seems that, on an average, about 5 percent of the Soviet capital stock is retired annually, giving an average life of the capital stock of twenty years.

The estimated system was used to produce a simulated run for the period 1950–75, with the exogenously specified values of (i) the 1950 capital stock and, (ii) labor, L_t, in billion man-hours during 1950–75 and the foreign resource inflow F_t during 1950–75. The simulation "tracked" the actual values of Soviet output rather well,[5] increasing our confidence in the exercises to which we turn next.

III. Estimating the Productivity of Foreign Resource Inflow

Given this estimated system, I simulated "runs" for the period 1976–85, using three alternative flows of foreign resources (F) at zero, 1 billion, and 3 billion in 1970 rubles for each year during the ten-year period. The incremental values of Soviet output for each year, between the zero and 1 billion and between the zero and 3 billion inflow runs, then represent the corresponding "benefits" from the streams of 1 and 3 billion rubles worth of foreign inflow, respectively.[6] Common to all the runs are the initial conditions (i.e., the 1976 beginning-of-the-year capital stock) and the assumed growth of the labor force at the rate of 1.2 percent per annum, the only difference being in the assumed level of F.

[4]The low estimate of σ is consistent with estimates of σ in econometric studies of Soviet industry alone. See Martin Weitzman and the author.

[5]The simulated output values exceed the actual outputs in later years, the excess falling in the range of 2–9 percent.

[6]Calculating the productivity of foreign inflow by reference to F_t at a historical trend value, as distinct from F_t at zero, makes little difference to the estimates presented because of its low average value around 0.3 billion 1970 rubles annually. Besides, the trend is difficult to estimate with statistical significance because of wide annual fluctuations. The runs at F_t worth 1 and 3 billion 1970 rubles seem realistic in view of the currently anticipated utilization of foreign resources by the Soviet Union.

Among the striking features of my predicted scenarios are:

1) A steady deceleration in both the rates of growth of output and investment with output growth declining from 3.8 percent in 1977 to about 3.4 percent in 1985 and investment growth decelerating to around 4 percent in 1985.[7]

2) A steady decline in the marginal productivity of capital from about 6.6 percent in 1976 to 3.5 percent in 1985. My depiction of the Soviet economy in terms of a *CES* production function with a low elasticity of substitution between capital and labor makes it a "labor shortage" economy with low returns on capital.

3) A miniscule impact of varying foreign resource inflows on output and also investment since the foreign resource inflow has been and will remain an insignificant fraction of the size of the Soviet economy in terms of its output and investment.

We may now turn to the calculation of the "internal rate of return," as an estimate of the productivity of foreign resource inflow to the Soviet economy. For this purpose, let us utilize the incremental income streams associated with the inflow streams of 1 and 3 billion 1970 rubles, respectively. Note that in view of the finite time horizon of ten years that is imposed on the benefit streams, I calculate and add to the benefit streams the incremental capital stock at the end of the period. In this way, I account for the income benefits beyond the ten-year period which are otherwise lost by the truncation of the time horizon to ten years.

This procedure for calculating the internal rate of return of the stream of foreign resource inflows leads to the following estimates: 1.88 percent for the 1 billion stream and 0.8 percent for the 3 billion stream. Since the production function is characterized by substantial diminishing returns, the estimated rate of return declines as the horizon is lengthened. Thus, for example, if I had calculated the internal rate of return with 1981 rather than 1985 as the terminal year, this procedure would have yielded the estimates of 4.41 percent corresponding to the 1 billion resource inflows and 1.68 percent corresponding to the 3 billion resource inflow.

IV. Concluding Observations

My estimates of the productivity of foreign resource inflows in the Soviet economy are remarkably low compared to the market terms for credits that typically obtain in the private capital markets of the West. They suggest the following thoughts.

First, they may explain why the Soviet Union is keen to get softer terms and conditions. While soft terms are better than hard terms, by definition, the Soviet keenness to get them may follow from the fact that hard, commercial terms may tend to result in counterproductive borrowing, given the low returns domestically.

Second, the estimates may also explain the Soviet emphasis on getting associated technology rather than pure capital inflows. Unless the importation of technology is at terms that capture for the sellers of technology the full returns to the Soviet Union from its utilization (which is quite improbable), the incremental benefits to Soviet income from such technological imports, financed with the capital inflows, would raise the net return from such "joint" capital cum technology inflows above the rates of return that we have calculated.

Third, and related to the preceding remark, my results may also suggest that the Soviet Union is likely to be very particular about the areas in which foreign capital cum technology is available and accepted. If the returns in terms of the growth of income from foreign resource inflow in general are not substantial, the net benefit would have to be substantial in terms of other objectives such as diversification of the economy towards consumer goods in toto, or some particular types of consumer goods for the commercial terms of such capital cum technology imports to be acceptable to the Soviet Union.

[7]Note that the official estimates of output and investment in the past four to five years also show a deceleration with the *actual* growth rates in each category being slightly higher than the simulations of this paper. Indeed, such a deceleration of the Soviet economy in recent years, which is likely to continue in the near future, has been emphasized as its dominant characteristic.

REFERENCES

A. Bergson, "Soviet Economic Prospects: Concluding Observations," in Yves Laulan, ed., *Prospects for Soviet Economic Growth in the 1970's*, Brussels 1971.

_____, "Soviet Economic Perspectives—Towards a New Growth Model," *Prob. Communism*, Mar. 1973, *22*, 1–9.

P. Desai, "The Production Function and Technical Change in Postwar Soviet Industry: A Reexamination," *Amer. Econ. Rev.*, June 1976, *66*, 372–81.

M. Feshbach and S. Rapawy, "Soviet Population and Manpower Trends and Policies," in Joint Economic Committee, U.S. Congress, *Soviet Economy in a New Perspective*, Washington, Oct. 1976.

Donald Green and Christopher Higgins, *SOVMOD I: A Macroeconometric Model of the Soviet Union*, New York 1977.

R. Greenslade, "The Real Gross National Product of the U.S.S.R., 1950–1975," in Joint Economic Committee, U.S. Congress, *Soviet Economy in a New Perspective*, Washington, Oct. 1976.

M. Weitzman, "Soviet Postwar Economic Growth and Capital-Labor Substitution," *Amer. Econ. Rev.*, Sept. 1970, *60*, 676–92.

Ministrestvo Vneshnei Torgovli SSSR, *Vneshnyaya Torgovlya SSSR, Statisticheskii Obzor*, Moscow, various years.

Tsentral'noe Statisticheskoe Upravlenie pri Sovete Ministrov SSSR, *Narodnoe Khoziaistvo S.S.S.R.*, (*Narkhoz*), Moscow, various years.

Some Systemic Factors Contributing to the Convertible Currency Shortages of Centrally Planned Economies

By FRANKLYN D. HOLZMAN*

The European centrally planned economies (*CPE*s) have sustained chronic hard currency deficits since East-West trade began to expand in earnest about fifteen years ago. While their outstanding hard currency debts almost doubled over 1975–76 as a result of an inability to adjust quickly to the Western recession—a previously unsuspected vulnerability—other systemic factors rooted in Stalinist central planning as practiced in the *CPE*s, have been responsible for the more secular balance-of-payments problems.[1] I refer to the wide use of direct controls to allocate intermediate products, the prevalence of "taut" or over full-employment planning, and irrational domestic pricing. These have several implications for economic performance which are relevant to the *CPE*s hard currency balances of payments.

First, the *CPE*s tend to produce relatively low quality manufactured products and have a marked inability to "sell" their products in Western markets. Inability to compete successfully is not due to price, but, to quote a Hungarian economist, Imre Vajda, to deficiencies in "performance, reliability, ... appearance, packing, delivery and credit terms, assembling facilities, after-sale services, advertising, selling itself ..., primarily factors other than price ..." (p. 53). This ineptness results largely from lack of competition—the fact that domestic products are "distributed by the plan" rather than "sold" and that quantitative goals take precedence over qualitative goals. Further, "taut planning" results in sellers' markets, additionally weakening managerial incentives to improve quality. Nor does competition play a significant role in intrabloc foreign trade. This trade is characterized by large state trading agreements, protected markets, and little or no direct contact between the producing enterprise in one nation and consuming enterprise in the other.

Second (and related) is the well-known relative weakness of socialist nations in innovation and technological change. This is due to the absence of "competition" just noted, to rewards for innovation which are inadequate to offset the risks or overcome inertia, and to the dysfunctional organization of *R&D* establishments and their relations to operating enterprises.

Third, the *CPE*s trade with each other and with the West at roughly world prices, even though these prices usually have no organic or consistent relationship to domestic prices. Their exchange rates serve as units of account but not as real prices. Their currencies are not only totally inconvertible into each other, they are also largely inconvertible into goods—so-called "commodity inconvertibility" (see the author, 1978). That is to say, foreign importers (exporters) are not allowed to compete freely with local enterprises for products (markets) because this would disrupt the plan. This significantly reduces short-run *ad hoc* exports—most exports have to be planned long in advance.

These factors lead to at least three causes of persistent hard currency shortages: 1) the

*Professor of economics, Tufts University and associate, Harvard Russian Research Center. Some of the ideas in this paper appeared earlier in my 1973 article. A much longer current version is available on request. I am indebted to Abram Bergson for incisive criticisms of two earlier drafts.

[1]Other than systemic factors may also be responsible. For example, the current availability of Western investments and credits on reasonable terms and the present willingness of the *CPE*s to entertain such relations with the West is one such factor. It should also be noted that the LDCs and some advanced industrial nations also have chronic balance of payments problems. However, I argue that the factors to be mentioned below are unique to the *CPE*s.

"saleability" illusion; 2) the "terms of trade" illusion; and 3) the "macro-balance" illusion.[2] To these I now turn.

I. The Saleability Illusion

East-West trade expanded rapidly over the past ten–fifteen years as a result of many factors—détente, reduction in Western controls and in Eastern strictures against trading with capitalists, search for new markets by the West and for quality products and new technology by the East, etc. A limiting factor in the growth of East-West trade has been inability of the East to compete successfully in manufactured products (Standard International Trade Classification (*SITC* 5–8)) for reasons mentioned earlier, inhibiting the expansion of their exports, hence imports. The importance of manufactured goods to *CPE* trade is highlighted by the facts that (before the rise in price of oil) almost three-fourths of intrabloc trade was in categories *SITC* 5–8. Yet the Council of Mutual Economic Assistance (*CMEA*) manufactured exports to the European Economic Community (*EEC*) and the European Free Trade Association (*EFTA*) averaged only about 30 percent of total exports, whereas corresponding imports were between 80–90 percent (see United Nations).

To the extent that the *CPE*s are aware of their selling difficulties, they can tailor planned imports to amounts which can be financed through exports. Saleability problems, in this case, reduce the level of trade or lead to planned foreign debts. Apparently, however, the planners also suffer from a saleability illusion which usually leads them to overestimate the amount of exports they can sell each year in the West. When *ex ante* export plans are not fulfilled, the *ex ante* balanced trade plan becomes an *ex post* deficit (with unplanned drawing down of reserves or unplanned credits) and/or import plans cannot be fulfilled. Support for the hypothesis that a saleability illusion exists can be found in the Eastern literature. For example, the eminent Hungarian economist, the late Sandor Ausch has stated: "In many cases . . . the extent of exports 'planned' by individual CMEA countries exceeds what the capitalist market in question is able to absorb . . ." (p. 109). Speaking on the same issue, the United Nations Economic Commission for Europe puts the problem as follows: ". . . East-European planners and exporters experience considerable difficulties in assessing their possibilities for sales of manufactured goods to the industrialized countries of western Europe *and also in selling their goods on the markets in question once decisions to export have been taken*" (p. 115, emphasis added).

One would think that the saleability illusion would eventually disappear. At least four reasons can be given which may explain its persistence. First, and most important, Western markets account for no more than about 10 percent of any Soviet bloc nation's sales (domestic and foreign), in the case of the *USSR* less than 1 percent. The planners think that their products are saleable in the West on the basis of the continuing experience that at least 90 percent or more of their products *are* saleable either at home or in intrabloc trade. This experience is undoubtedly so overwhelming that it is difficult for the planners and especially plant managers to adapt in practice fully to the idiosyncracies (to them) of Western markets. Second, some attempts are undoubtedly made by bloc producers and salesmen to improve the "quality" and saleability of their products. But the quality, technology, etc. "gaps" are probably perceived by bloc planners as stationary goals when in fact they are moving targets, hence the gap remains. Third, many bloc foreign trade organization representatives undoubt-

[2]While not precisely comparable, parallels with Western experience may be drawn for illustrative purposes. A Communist country with a "saleability" illusion and very price inelastic demands for its exports is one which faces balance-of-payments problems similar to those faced by a capitalist nation in "structural" disequilibrium. Nations which run deficits because their exportables have recently become obsolete or have been exhausted (raw materials) are cases in point. The "terms of trade" illusion is experienced by Western nations with overvalued exchange rates. Western nations with full employment and inflation are apt to experience balance-of-payments problems related to those of communist nations with what I have called a macro-balance illusion.

edly do realize that the products they are trying to sell are deficient by Western standards. It is one thing to recognize the problem; it is another, however, to force producing enterprises, advertisers, packagers, servicers, etc. to meet higher specifications when they have little or no motivation to do so. Finally, the foreign trade plans may be chronically too "taut" just as the national domestic economic plans are and for analogous reasons. Possessed of a great desire for imports from the West, the planners attempt to sell more to the West than is feasible; and the more they wish to buy from the West, the greater the temptation to try to export products which may be unsaleable.

No analogous market impediments exist on the import side. Imports are consummated according to plan. When export earnings fall behind, a deficit results and/or imports are repressed. It is hard to immediately repress imports, especially of intermediate products, because of negative repercussions on domestic plan fulfillment. Witness the lag in reducing imports during the Western recession of 1974–76.

To sum up: the *CMEA* nations establish plans envisaging earnings from exports sufficient to finance a planned quantum of imports. Actual exports fall short of planned exports resulting in an excess demand which may be satisfied either through borrowing or drawing down reserves. The rapid rise in hard currency debt of *CMEA* is evidence of borrowing and some of this was undoubtedly "unplanned." The unsold exportables are not likely to be observable, however, since they are rapidly absorbed into the domestic economy when proved unsaleable on world markets (see Section III).

II. Inability to Devalue and the Terms of Trade Illusion: An Import/Export Asymmetry

Bloc currencies cannot be viewed as overvalued in the usual Western sense. Nevertheless the foreign trade behavior described in Section I, which results in either unplanned deficits or the need to suppress planned imports, does suggest the equivalent of overvaluation.

Another factor which suggests the existence of (the equivalent of) overvaluation is the fact that the Socialist nations can be viewed as a high-cost, low-variety, low-quality economic region relative to the rest of the world. They have constituted, in effect, the equivalent of a trade-diverting customs union. It follows that any relaxation of controls or mutual reduction of East-West barriers will lead to a tendency toward more imports by East than by West.

While under capitalist institutions, the above circumstances would spell currency overvaluation and would call for devaluation (or downward float) as a remedy, under present Socialist institutions the same cannot be said. As noted, their currencies do not function as means of payment and their official exchange rates do not serve as real prices (exclude Hungary). Use of world prices and convertible currencies does effectively circumvent the need to use Socialist currencies and exchange rates, but by the same token it deprives these nations of an important instrument variable for improving their balances of payments, namely, devaluation. Unable to devalue, the *CPE*s' foreign trade planners trade with the West at sets of world prices which make them deficit prone, as are Western nations with overvalued currencies. Under capitalism, maintenance of an overvalued exchange rate leads to deficits because the individual importers and exporters receive misleading signals—prices including hidden subsidies and tariffs, respectively—and the capitalist nation, taken as a whole, operates under a "term of trade" illusion; devaluation eliminates implicit subsidies and tariffs, and restores payments equilibrium. The *CPE* planners cannot devalue to eliminate a deficit. Is it possible to simulate devaluation? The answer is yes for exports—but it is not clear that such a policy will be blessed with success.

Devaluation can be simulated on the export side by simply reducing below (or further below) world prices the prices which exports are offered. Pricing exports below world prices may not always be feasible because of antidumping or antimarket disruption rules which typically cannot be refuted. Neverthe-

less, suppose export prices can be lowered—what will the market response be? If demand is elastic, export earnings can be increased financing more imports—but not as much as planned because of the deterioration in terms of trade. *In the short run, this implies borrowing* and also perhaps some cutting back of imports. But Western demand may be very price inelastic for most Eastern products (excluding Soviet raw materials). To paraphrase Vajda—lack of competitiveness is mainly due to nonprice factors. Western buyers don't want obsolete, low-quality, etc. products at almost any price. Potentially immiserizing terms of trade lead to borrowing even more in the short run to fulfill import plans, with cutbacks in imports and attempts to develop exportables over the longer run.

The situation described above (and also in Section I) is a "bind" akin to that of Western nations that have experienced so-called "structural disequilibrium" in their balances of payments (see Charles Kindleberger, pp. 487–88). This term has been used to describe nations whose exportables have been, among other things, 1) products which have become obsolete, in some sense, and can no longer compete in world markets or 2) resources which have been exhausted or 3) temporarily reduced by wartime destruction and disruption (Europe after World War II). At the same time, nations in the first two categories have become *accustomed* to a certain level of imports either as crucial intermediate inputs into industry or as final products in the standard of living or, if in the third category, urgently *needed* imports for reconstruction. Under these circumstances, devaluation is not likely to increase exports in the short or medium run—not until substitute exportables can be developed. Devaluation would, sharply reduce expenditures on imports if demand were sufficiently elastic—which under these conditions may not always be the case. Hence strong balance-of-payments pressures may exist for some time.

I have discussed the simulation of devaluation on the export side, but not with regard to imports. Simulation on the import side is more difficult since if it is to be accomplished it must be done so in terms of shadow prices rather than real prices. Let me explain. Devaluation raises the actual prices of imports to buyers in the devaluing nation and this serves as a strong disincentive to importing as much as before. This cannot be simulated by a Socialist nation—that is to say, no nation would insist on paying more than the going price for imports. What the nation can do, however, is to raise the minimum level of profitability at which imports are allowed.[3] Such an effort is likely to be less than perfectly satisfactory. With prices as messed up as they are, profitability measures are unlikely to be taken very seriously.[4] It will be difficult, indeed, to reduce the level of desired imports when their ostensible hard currency cost has not changed. An overvaluation-type illusion that imports are cheaper than they really are undoubtedly remains under these circumstances. Further, the antimercantilist approach of Communist planners gives them a tendency to be overzealous importers and underzealous exporters particularly in the absence of clear terms of trade signals.

III. The Macro-Balance Illusion

All the *CPE*s practice overfull-employment planning which means that planned demands exceed available supplies. Under these circumstances, as with inflationary pressures in Western nations, domestic producers and consumers compete for exportables and

[3]Actually, many of the *CPE*s are reported to use foreign trade effectiveness indexes as a guide regarding what to import and export. In their simplest form, these are ratios of local currency prices of exports or import substitutes over foreign trade prices in foreign currencies. They tell the planners how much in domestic resources is required to earn a dollar of foreign exchange through exports of different products; and how much in domestic resources are saved by a dollar's worth of imports of different products. It should be profitable to export commodities with low ratios and to import commodities *with high ratios. Simulation of devaluation involves*, in part, raising the maximum ratio at which exports are promoted and raising the minimum ratio at which imports are allowed.

[4]Discussions with a number of Eastern European and Soviet foreign trade specialists in June 1976 confirmed to me that in most of the *CPE*s, these indexes are either not used extensively or provide only one of many kinds of information upon which import and export decisions are based.

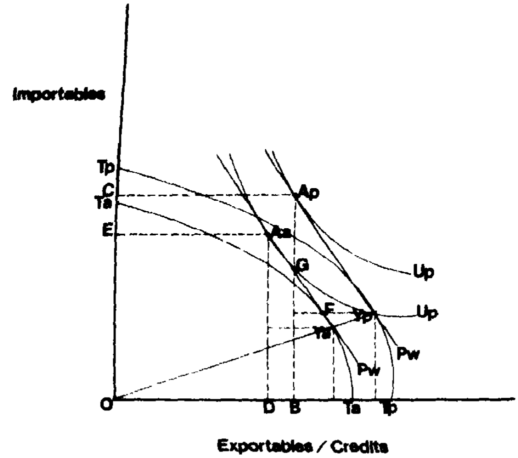


FIGURE 1

demand more imports, and in the process create pressures which, if successful, cause deterioration in the balance of payments, including that with the West. These forces are most easily envisaged in terms of the "absorption" approach:

$$Y - A = X - M$$

where X and M denote exports and imports, respectively; Y denotes output and A is absorption or expenditures (for $C + I + G$). Clearly, if a nation spends more than it produces, i.e., $A > Y$, then it must run a deficit, $M > X$.

The situation can be demonstrated diagrammatically as follows (see Figure 1). Assume a nation on an actual transformation curve T_a, with domestic output at Y_a, and an after-trade equivalent absorption A_a. Assume that this nation planned to be on T_p with before-trade goal of Y_p and after-trade goal of A_p. If the nation actually ends up at A_a instead of planned A_p it will experience excess demand equal to BD of exportables + CE of importables (or some equivalent combination at some other point such as A_pG of importables). Plans may be fulfilled by either dishoarding foreign exchange reserves or obtaining credits of an amount measured as FY_p of exportables. An example of this is illustrated by the Polish economist Stanislav Gruzevskii writing in 1968 regarding the situation in Poland in 1966–67. He states that industry, particularly the machine building industry "... did not have at its disposal the necessary production capacity and therefore was unable to meet its obligations for exports within the group of commodities including machinery and equipment, and which also could not meet the demand of the domestic market. The result of the latter failure was that the plan of imports of machinery and equipment for the same period was exceeded ..." (p. 22).

REFERENCES

Sandor Ausch, *The Theory and Practice of CMEA Cooperation*, Budapest 1972.

S. Gruzevskii, "Direction of Export Specialization in the Machine-Building Industry," *Sov. East Euro. Foreign Trade*, No. 4, 1968–69, 21–34.

F. D. Holzman, "East-West Trade and Investment Policy Issues," in Joint Economic Committee, U.S. Congress, *Soviet Economic Prospects for the Seventies*, Washington 1973, 686–89.

———, "Ruble Convertibility," in Nita Watts, ed., *Economic Relations Between East and West*, New York; London 1978.

Charles P. Kindleberger, *International Economics*, 4th ed., Homewood 1968.

I. Vajda, "External Equilibrium and Economic Reform" in Imre Vajda and Michael Simai, eds., *Foreign Trade in a Planned Economy*, Cambridge 1971.

United Nations, Economic Commission for Europe, *Analytical Report on the State of Intra-European Trade*, New York 1970.

THE EFFECTIVENESS OF FISCAL POLICY

Temporary Taxes as Macro-Economic Stabilizers

By WALTER DOLDE*

An analysis of the effectiveness of temporary tax changes requires both a theoretical framework and its careful empirical implementation. Reflex rejection of the usefulness of temporary taxes with a vague appeal to the permanent income hypothesis is as inappropriate as blind acceptance naively based on the high correlation between consumer expenditure and current disposable income. The remainder of this paper sketches a theoretical framework for analysis with an eye towards implementation, reviews the evidence for the United States, and provides a summary. The analysis concludes that temporary taxes are useful and effective stabilization instruments, though there is no reason to favor them over tax changes of an indefinite duration. Space limitations prevent discussion of taxes other than personal income taxes. Expenditure taxes, for which intertemporal substitution effects augment income effects on current expenditure, have a greater effect per dollar of deficit.

I. Theoretical Framework

The important questions in investigating temporary taxes relate to timing. A change in household budget constraints must be spent sometime, but spending changes occurring after more than six or eight quarters probably are of little help for stabilization purposes. The life cycle hypothesis (*LCH*) and the permanent income hypothesis (*PIH*) provide a natural starting point for analyzing consumer behavior in a dynamic setting.

Consider the model in equation (1):

$$(1) \qquad c_t^a = b(d^a,a)\left[w_t^a + \sum_s d_{t,s}^a y_{t+s}^a\right]$$

A household is indexed by its age, a. Consumption in year t is proportional to the sum of accumulated property net worth (w_t^a) and the present value of expected remaining nonproperty income net of tax liabilities. The applicable discount factor for the cash flows s years hence is $d_{t,s}^a$. The proportionality factor b, the average (*APC*) and marginal (*MPC*) propensity to consume from life cycle resources, depends on demographic variables summarized by a and on the whole vector of discount rates d^a.

Three distinct dynamic elements affect consumer responses to temporary taxes: 1) consumer beliefs about the permanence of tax changes, represented by changes in expected future income (y_{t+s}^a), 2) discount rates (d^a), and 3) marginal propensities to spend now (b) and in the future.

Factors which cause current income to change may also cause changes in expected future income (y_{t+s}^a). The higher the *permanence presumption ratio*[1] (*PPR*) about an income change, the greater is consumer response. Obviously lower *PPR*s are associated with temporary than with permanent tax changes. It is less obvious that *PPRs* for temporary tax changes are lower than those for most income fluctuations. Tax legislation is much discussed in the media and on the street. Consumers learn about it very rapidly. Consumers learn about and infer the permanence of many other income fluctuations only over time, as in the framework discussed by John Muth. This argument applies to tempo-

*Graduate School of Industrial Administration, Carnegie-Mellon University. I am grateful to Robert J. Hodrick, Lester B. Lave, and Edward C. Prescott for comments and criticisms. Remaining errors are my own. Research support from the National Science Foundation is appreciatively acknowledged.

[1]This term was suggested by Arthur Okun in personal correspondence.

rary tax changes lasting a few quarters but not, of course, to rebates.

The credibility of the government in consumers' eyes also affects the *PPR* associated with temporary tax changes. The cynical observer may doubt that tax increases announced as temporary will actually be so. Indeed the 1968 surcharge was extended beyond the termination date announced at its inception. Symmetry suggests that the *PPR* associated with temporary tax cuts might thus be moderated. In fact, the 1975 tax cut was subsequently extended. The need to offset fiscal drag in a growing economy is one factor that would cause temporary tax cuts to be extended.

Finally it is worth noting that many "permanent" tax changes have not lasted much longer than those announced as temporary. There were eleven major federal tax change acts between 1945 and 1975, according to the Tax Foundation, Inc. (p. 103). The year 1977 brought a twelfth, and Congress is considering a thirteenth at present. The same source (p. 34) provides another measure of the frequency of permanent tax changes. In fourteen out of twenty-nine years the percentage of net national product collected as tax receipts at all government levels changed in absolute value by more than 1 percent (of net national product).

The second factor prominent in the *LCH-PIH* framework and important here is the rate at which consumers discount future income in estimating the value of life cycle resources. Milton Friedman, Thomas Juster, and the author (1978) have suggested that on average consumers use discount rates of 33-1/3–40 percent and higher ($d^a_{t,t+1} \approx .6$ or lower). The illiquidity of the main asset owned by most families, human capital, helps to explain these high discount rates. Over its life cycle the annual labor income of a generation will quintuple in constant dollars (see the author, 1978). Yet the ability of most households to borrow more than trivial amounts against this future nonproperty income is nonexistent. Alan Blinder (1976) has emphasized that far from being restricted to households with low life cycle income profiles, these liquidity constraints are likely to bind more severely on households with high-income profiles. High-income profiles tend to be more skewed to the end of the life cycle than low-income profiles. Households with high-income profiles are more likely to anticipate a nontrivial inheritance, which is also quite illiquid.

Liquidity constraints make the shadow price of consuming future income today very high and make the values of $d^a_{t,s}$ very low. At low discount factors, current and near future income flows receive much higher relative weights in the estimate of permanent income or the value today of life cycle resources. Changes in income flows occurring more than a few years in the future are irrelevant to current spending: $d^a_{t,s} \approx 0$ for $s > 4$ or 5 years. At an annual discount rate of 40 percent, a two-year tax change has 49 percent as great an effect on the estimate of permanent or life cycle income as does a tax change that lasts literally forever. A two-year tax change would have 66 percent of the effect on permanent income that a four-year tax change would have. To put this another way, at low discount factors the equivalent (literally) permanent tax changes must be on the order of half or more as large as temporary tax changes.

The *MPC* for the current period b comprises the third important intertemporal element in evaluating temporary tax effectiveness. One of the important testable propositions of the *LCH* is that b increases with the age of the household, $\partial b/\partial a > 0$. For an individual household, b is independent of the source—labor income, property income, transfer payments, tax liabilities—of increments to life cycle resources. (An exception is property income changes accompanied by changes in d^a.) Aggregate *MPC*s differ, however, because changes in various types of income and tax liabilities are distributed across different subpopulations of households. A higher aggregate *MPC* should be associated with an increase in transfer payments to the retired than with an increase in corporate dividend payments.

Consideration of age alone suggests that general tax changes should meet with a lower aggregate *MPC* than do variations in other types of income. The share in tax liabilities of the working age population is greater than its

proportion of aggregate disposable income. A more careful examination of demographic factors, however, modifies the simpler conclusion. While b should be monotonically increasing in a for single person households it may not be monotonic for multiperson households. In particular b may be very high during the child rearing years. Thus while tax-caused changes in life cycle resources are distributed across households with a number of years left in their life cycles, some of those households have substantial *MPC*s for the current year because family size is near its life cycle peak.

The *MPC* depends on the discount rates in d^a. Near-term liquidity constraints shorten the period over which consumption smoothing occurs and raise the *MPC* for the current period. In the limit of currently binding liquidity constraints, the value of life cycle resources for current consumption is identical to those resources currently available and b takes on the value unity (see the author, 1978, p. 418).

In summarizing a theoretical framework for investigating the effectiveness of temporary taxes, it is easiest to describe what will not work. What will not work is a regression of consumption on current and lagged aggregate income. Income must be disaggregated by type since the aggregate *MPC* from tax liabilities may be different from aggregate *MPC*s for other income types. Great care must be given to expectations since consumers may learn about tax changes more rapidly and may attribute a different permanence to them than to other income changes. This problem may not be susceptible to mechanical estimation techniques. The economic analyst may have to substitute his own range of forecasts of consumers' views of the permanence of tax changes. The effects of this difficulty are mitigated by the high rates at which households discount the future on average.

II. Evidence

The theoretical discussion above ignored distinctions between consumer spending and consumption. While the latter comprises the arguments of utility functions, it is the former which is of concern for macro-economic stabilization. Influence on consumer spending provides the measure by which the effectiveness of temporary taxes should be judged. Many who concluded the 1968 surcharge was ineffective pointed to its inability to reduce inflation. The great recession of the 1970's has demonstrated that the insensitivity of inflation to major reductions in aggregate demand below capacity was not specific to the 1968–70 period.

The 1968 tax surcharge provided a challenge for econometric inference. The consumption effects of normal income fluctuations can be inferred from an econometric model by forecasting consumption both with and without the income changes. Call these forecasts C^w and C^{wo}, respectively. Then $C^w - C^{wo}$ is the implied effect. This technique may be inappropriate for inferring the consumption effects of a temporary tax change, however, if the corresponding income fluctuation is unlike the income fluctuations that occurred in the historical sample period (see Robert Lucas).

Okun (1971) proposed the alternative of comparing actual consumption C with the bounds provided by C^w and C^{wo} as forecasted by four well-known econometric models. Comparing C with C^w corresponds to what Okun calls the *full effect view*: the tax surcharge had the full consumption effect of any other income change. The comparison of C and C^{wo} results in the *zero effect view:* the tax surcharge was ignored by consumers. Theory does not lead to the zero effect view, of course, except in the case of an infinitely lived individual, and then only if he has zero time preference. Okun found the full effect view to result in smaller root-mean squared errors (*RMSE*) over the surcharge period both for nondurables and services and for durables excluding autos. The finding on nondurables and services is especially significant since no expenditure vis-à-vis consumption of services problems cloud the inference in this case. For the final component of consumer expenditures, automobile purchases, the zero effect view had a lower *RMSE*. Even the zero effect view seriously *underpredicted* automobile purchases during the surcharge period, however, suggesting an

autonomous increase in automobile demand unrelated to the surcharge.

Blinder and Robert Solow extended Okun's technique to linear combinations intermediate between the bounds provided by the zero effect and full effect views. They find the lowest *RMSE*s result when the surcharge is assumed to have been 50 percent as effective as any other income change.

William Springer (1975, 1977) and Okun (1977) disagreed about statistical technique and about the effectiveness of the surcharge. The equations used by Okun generally have smaller standard errors and *RMSE*s than Springer's estimated equation.

Franco Modigliani and Charles Steindel, using the same technique with the MIT-Penn-SSRC model, also judged the 1968–70 surcharge to have been about half as effective as a permanent tax change. Modigliani and Steindel performed the same analysis on the *rebate* portion of the 1975 tax changes, treating the *temporary* aspects like any other income change. They found the rebate to have been about half as effective through its first three quarters (1975 II–IV) as a permanent tax cut followed in the next quarter by an equivalent permanent tax increase. They infer that its effectiveness was even greater in the succeeding three quarters.

Blinder (1977) separated the temporary tax changes of 1968 and 1975 from other income and estimated the differential responses to windfalls, permanent income changes, and to the temporary tax changes. Depending on the treatment of rebates, Blinder finds temporary taxes are 66 percent as effective as other income changes in their first quarter, and 77–92 percent as effective within a year. Rebates are 55–66 percent as effective initially and 57–77 percent as effective after a year.

While he did not directly address the issue of temporary taxes, Michael Darby's work (1972, 1974) is relevant here because he finds average consumer discount rates to be quite low; about 10 percent ($1 - d^{aw}_{t,t+1} \approx .1$). In Darby's empirical framework, however, the discount parameter serves a dual role relating both to discounting and to forming *PPR*s adaptively as suggested by Muth. Darby (1972, p. 938) finds that changes in current income evoke *MPC*s of .55 to .74 initially. These *MPC*s rise over time for maintained income changes.

Finally, one might take the view that consumers respond much the same to temporary and permanent tax changes, given the frequent occurrence of the latter. Aggregate responses to tax changes need not, however, be identical to aggregate responses to other income changes. In previous work (1976), I found that consumption of nondurables and services responds more rapidly to tax changes than to other income changes. The estimated average discount rate associated with before-tax income was 15 percent per quarter ($d^{aw}_{t,t+25} = .87$), in agreement with the discussion above. The estimated average discount rate associated with tax liabilities was 123 percent per quarter ($d^{aw}_{t,t+25} = .45$). The lags for forming *PPR*s also differed. The mean lag associated with tax liabilities was 1.2 quarters, shorter than the 2.2 quarters for other income changes.

III. Conclusions

An infinitely lived individual with unlimited borrowing opportunities has the opportunity to insulate his consumption from temporary tax changes. With appropriate preferences, he might actually choose to do so. The circumstances of *U.S.* households differ greatly from this abstraction, and so does their expenditure behavior in the face of temporary taxes. The bulk of their life cycle resources takes the form of illiquid future nonproperty income which they apparently discount at annual rates of 40 percent and more. Households learn about the permanence of tax changes, because of the attendant publicity, more rapidly than they learn about other income changes. As a consequence they adjust their expenditures more rapidly after tax changes. The temporary tax changes of 1968 and 1975 were only two of a dozen major federal tax changes since 1945. Thus rational consumer responses to temporary tax changes are only slightly smaller than responses to other tax changes.

The frequency of nontemporary tax

changes also mitigates Lucas' criticism of econometric policy evaluation in this case. Further, it is possible to estimate bounds on the relative effectiveness of temporary tax changes. Most of the evidence points to a relative effectiveness on consumer expenditures of 50 to 90 percent of permanent tax changes.

Political considerations provide the most plausible explanation for the past use of temporary taxes. On economic grounds, there is little reason to favor either temporary tax changes or tax changes of an indefinite duration. A permanent tax equivalent exists for any temporary tax, which is heavily front loaded by comparison. Because of these discount rates and other factors the permanent tax equivalent, far from being an order of magnitude smaller, would be between 50 and 90 percent as large as a temporary change.

REFERENCES

A. S. Blinder, "Intergenerational Transfers and Life Cycle Consumption," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 87–93.

——, "Temporary Taxes and Consumer Spending," unpublished paper, Princeton Univ., Apr. 1977.

—— **and R. M. Solow,** "Analytical Foundations of Fiscal Policy," in Alan S. Blinder et al., eds., *The Economics of Public Finance*, Washington 1974.

M. R. Darby, "The Allocation of Transitory Income Among Consumers' Assets," *Amer. Econ. Rev.*, Dec. 1972, *62*, 928–41.

——, "The Permanent Income Theory of Consumption—A Restatement," *Quart. J. Econ.*, May 1974, *88*, 228–50.

W. Dolde, "Forecasting the Consumption Effects of Stabilization Policies," *Int. Econ. Rev.*, June 1976, *17*, 431–46.

——, "Capital Markets and the Short Run Behavior of Life Cycle Savers," *J. Finance*, May 1978, *33*, 413–28.

M. Friedman, "Windfalls, the 'Horizon,' and Related Concepts in the Permanent-Income Hypothesis," in Carl F. Christ et al., eds., *Measurement in Economics*, Stanford 1963.

T. F. Juster, "Discussion," in *Consumer Spending and Monetary Policy: The Linkages*, Federal Reserve Bank of Boston, Boston 1971.

R. E. Lucas, Jr., "Econometric Policy Evaluation: A Critique," *J. Monet. Econ.*, Jan. 1976 suppl., *2*, 19–46.

F. Modigliani and C. Steindel, "Is a Tax Rebate an Effective Tool for Stabilization Policy?," *Brookings Papers*, Washington 1977, *1*, 175-209.

J. F. Muth, "Optimal Properties of Exponentially Weighted Forecasts," *J. Amer. Statist. Assn.*, June 1960, *55*, 299–306.

A. M. Okun, "The Personal Tax Surcharge and Consumer Demand, 1968–70," *Brookings Papers*, Washington 1971, *1*, 167–211.

——, "Did the Surcharge Really Work?: Comment," *Amer. Econ. Rev.*, Mar. 1977, *67*, 166–69.

W. L. Springer, "Did the 1968 Surcharge Really Work?," *Amer. Econ. Rev.*, Sept. 1975, *65*, 644–59.

——, "Did the 1968 Surcharge Really Work?: Reply," *Amer. Econ. Rev.*, Mar. 1977, *67*, 170–72.

Tax Foundation, Inc., *Facts and Figures on Government Finance*, New York 1977.

On Modeling the Effects of Government Policies

By RAY C. FAIR*

An important question in macroeconomics is how government policies affect the economy. The fact that this question is still being debated forty-three years after John Maynard Keynes wrote *The General Theory* and thirty-nine years after Jan Tinbergen did his pioneering econometric study of business cycles for the League of Nations indicates the difficulty of answering it. Although it is easy to construct theoretical models in which government policies do or do not have important effects on, say, real output, it is difficult to test alternative models. One difficulty is the relative ease with which aggregate time-series data can be fit well within the sample period. Because of this, a good within-sample fit is by no means a guarantee that the particular equation or model is a good representation of the actual process generating the data. It is also difficult to make comparisons of predictive accuracy across models because of differences in the number and types of variables that are taken to be exogenous. These difficulties are annoying and have undoubtedly contributed to giving macroeconomics a bad name.

The purpose of this paper is to review that part of my recent work that relates to modeling the effects of government policies. In my econometric model, government actions, even if they are anticipated, can have important effects on real output, and in Section I the theoretical basis for this property is reviewed. In Section II the sensitivity of policy effects in the model to a number of alternative assumptions is examined. These assumptions concern 1) the behavior of the Federal Reserve, 2) whether or not there are rational expectations in the bond and stock markets, and 3) whether or not government bonds are treated as wealth by the household sector. In Section III the testing of the model is discussed. Two types of tests are considered in this section: tests of individual hypotheses and tests of the accuracy of the overall model. The main points of the paper are summarized in Section IV.

I. Theoretical Issues

The proposition that government tax rates and transfer payments affect the decisions of households and firms is familiar from microeconomics. In the theoretical model upon which my econometric model is based (see the author, 1974, 1976), the "micro-economic" aspect of individual decisions has been stressed. The decisions of the individual agents in the model (households, firms, and banks) are derived from the solutions of multiperiod optimization problems. At the beginning of each period each agent solves its optimization problem, knowing all past values, receiving in some cases information from others regarding certain current-period values, and forming expectations of future values. A number of government policy variables affect the solutions of these problems, and so through this channel government actions affect the economy. Tax rates and transfer payments, for example, affect the labor-leisure choice of the utility-maximizing households.

Although these micro-economic effects are fairly well accepted in the profession, they do not exist in a popular class of rational expectations macro models (see Robert J. Barro, Robert E. Lucas, Jr., Thomas J. Sargent, 1973, 1976, and Sargent and Neil Wallace). Elsewhere (1978b), I have criticized this class of models for postulating that individuals are rational with respect to their expectation formation but not rational with respect to their overall behavior. This is an important criticism of these models in that their key property regarding the ineffectiveness of anticipated government actions on real output

*Cowles Foundation, department of economics, Yale University. The research described in this paper was financed by grant SOC77-03274 from the National Science Foundation.

no longer holds if rationality with respect to overall behavior is introduced into the models. In a "completely" rational model the government can affect real output by affecting, among other things, the labor-leisure choice of households.

There is also in my theoretical model another reason government actions can affect the economy. The model allows for the existence of disequilibrium, and if there is disequilibrium, the government can, by conventional means, help to correct it. Disequilibrium takes the form of banks constraining firms and households in how much money they can borrow at the current loan rates and of firms constraining households in how much they can work at the current wage rates. Binding constraints in the loan market are due to mistakes on the part of banks in setting loan rates, and binding constraints in the labor market are due to mistakes on the part of firms in setting prices and wages. These mistakes are the result of expectation errors. No agent knows the complete model, and so expectations can turn out to be wrong even though there are no random shocks in the model. There is, however, a continual adjustment to past mistakes in that each period the individual agents form a new set of expectations and reoptimize on the basis of information from the previous period.

The key premise of my theoretical work is thus that agents each period first form a set of expectations of future values and then given these expectations base their decisions on the solutions of multiperiod optimization problems. The expectations may be in error, and so banks and firms may set values of loan rates, wages, and prices that are not market clearing. Both the micro-economic and disequilibrium aspects of this premise imply that the government can affect real variables in the economy.

One important difference between my theoretical model and a disequilibrium model like that of Barro and Herschel Grossman should be noted. In the Barro-Grossman model, prices and wages are not decision variables of firms (or any other agents), and no explanation is provided as to why it is that prices and wages may not always clear markets. In my model, on the other hand, such an explanation is provided, namely the possibility of expectation errors on the part of firms. In short, my model, unlike Barro and Grossman's, is "choice theoretic" with respect to the determination of prices and wages. Because of this weakness of the Barro-Grossman model, Grossman is now advocating another theory of employment fluctuations, a theory in which market transactions are viewed as involving implicit contractual arrangements for mitigating risk.

II. The Sensitivity of Policy Effects to Alternative Assumptions

The properties of macro-econometric models tend to be sensitive to alternative assumptions. Given the difficulty of testing assumptions, this means that any policy recommendations that result from analyzing a model must be interpreted with considerable caution. Some assumptions are, however, more important than others in this regard, and the purpose of this section is to review the sensitivity results that I have obtained with my model.

Fiscal policy effects in the model are, as reported in my 1978a paper, quite sensitive to assumptions about monetary policy. The results of five experiments are presented here. Each experiment corresponded to the same fiscal policy shock (an increase in government purchases of goods). For four of the experiments the behavior of the Fed was assumed to be exogenous: in each of these cases the Fed was assumed to control a particular variable, which was then taken to be exogenous for the experiment. (By exogenous here is meant that for the experiment the variable was kept unchanged each period from its base-simulation value.) The control variables in the four cases were: 1) the amount of government securities outstanding; 2) the money supply; 3) nonborrowed reserves; and 4) the bill rate. For the fifth experiment the Fed was assumed to behave according to an estimated equation. The behavior that is reflected in this equation is behavior in which the Fed "leans against the wind." As the economy expands or as

inflation increases, the Fed is estimated to cause interest rates to rise.

The results of these experiments are briefly as follows. When the Fed behaved according to the estimated equation, the sum of the increase in real output over the first twelve quarters after the fiscal policy change was 61 percent of the sum when the bill rate was kept unchanged. When the money supply was kept unchanged, the sum was 45 percent of the sum in the constant bill-rate case. The most expansionary case was the one in which the amount of government securities outstanding was kept unchanged. In this case the government deficit that results from the fiscal policy change is financed by an increase in high powered money. The sum in this case was 107 percent of the sum in the constant bill-rate case. Finally, the sum in the case in which nonborrowed reserves was kept unchanged was 72 percent of the sum in the constant bill-rate case. In short, these results indicate that fiscal policy effects are quite sensitive to what is assumed about Fed behavior.

Policy effects in the model are also sensitive to what is assumed about expectations in the bond and stock markets. In my forthcoming paper I have examined policy effects in three versions of the model: 1) the regular version (Model 1), in which expectations of future interest rates and stock prices are not rational; 2) a version (Model 2) in which expectations of future interest rates are rational; and 3) a version (Model 3) in which expectations of future interest rates and stock prices are rational. The fiscal policy shock described above was used, and the Fed was assumed to behave according to the estimated equation mentioned above.

The results for these three versions of the model are as follows. For Model 2 the sum of the increase in real output over the first twelve quarters after the fiscal policy change was 57 percent of the sum for Model 1. In Model 2 people know that the Fed is going to respond to the fiscal policy stimulus by increasing interest rates in the future, and this information gets incorporated immediately into long-term rates. In Model 1, on the other hand, long-term rates adjust only to the current and lagged increases in the short rate. Higher long-term rates have, other things being equal, a contractionary effect on the economy, and this is the main reason for the smaller increases in real output in Model 2 than in Model 1. For Model 3 the sum of the output increase was 61 percent of the sum for Model 1. In Model 3 people also know that profits are going to be higher in the future as a result of the stimulus, and this information gets incorporated immediately into stock prices. Stock prices are thus higher in Model 3 than they are in Model 2, and this leads, through a wealth effect on the household sector, to a slightly more expansionary economy in Model 3 than in Model 2. This difference is, however, much smaller than the difference between the results for Models 1 and 2, and so in this sense expected future profits in the model are less important than expected future interest rates.

The experiment just described was an unanticipated fiscal policy change. I also ran an experiment in which the change was announced thirteen years before it was actually made. In this case in Models 2 and 3 (but not in Model 1) people begin to adjust to the higher expected future interest rates and profits before the change is actually made. For Model 3 the sum of the change in output between the time of the announcement and twelve quarters after the change was actually made was 29 percent of the sum for Model 1. The anticipated policy change was thus about half as stimulative as the unanticipated change (29 vs. 61 percent).

Although the results just described are clearly tentative and are in no way a test of the assumption of rational expectations, they do indicate that this assumption is of considerably quantitative significance in macroeconometric models.

For purposes of the present paper I have also examined the sensitivity of policy effects to the treatment of government debt as wealth by the household sector. The wealth of the household sector is an explanatory variable in the four consumption equations in the model and in one of the three labor supply equations. In the regular version of the model government debt is included in this wealth variable. For an alternative version I reestimated the

five equations with government debt subtracted from the wealth variable. I then applied the same fiscal policy shock described above to this version (with the Fed behaving according to the estimated equation). The results from this exercise are easy to summarize. First, the fits of the five equations in the alternative version were almost identical to the corresponding fits in the regular version. It is clear that the macro data are not adequate for discriminating between these two versions of the model. Second, and for once fortunately, the policy properties of the two versions are quite similar. In other words, policy effects in the model are not sensitive to whether or not government debt is treated as wealth by the household sector. The sum of the increase in output over the twelve quarters in the alternative version was 98 percent of the sum in the regular version.

III. Tests of the Model

Given that my model does allow for anticipated government actions to have effects on real output, it is of considerable interest to test this model against models that do not allow for these effects. In this section the tests of the model that I have performed will be reviewed.

There are, first of all, t-tests of the individual coefficient estimates. A number of the explanatory variables in the consumption and labor supply equations that one expects from micro-economic theory to affect the decisions of households are significant by conventional standards. In addition, the "disequilibrium" variable that I have used to try to account for possible work constraints on the household sector is significant in a number of the equations. One must, of course, be skeptical of t-tests because of the ease with which good t-values can be obtained in macro data and of the general problem of data mining, and the purported significance of my micro-economic and disequilibrium explanatory variables is no exception to this. At least with respect to the disequilibrium variable, however, it does seem unlikely to me that the results would be as they are if there were no disequilibrium effects in the economy. The variable (ZJ) appears in the four consumption equations and in two of the three labor supply equations, with t-values for the most recent set of estimates, (see my 1978e paper), of 1.49, 3.37, 4.43, 2.17, 3.92, and 2.36.

With respect to possible tests of the assumptions examined in the previous section, I have already mentioned that it seems unlikely that the data are adequate for testing the hypothesis that government debt is treated as wealth by households. It may, on the other hand, be possible to test the hypothesis of rational expectations in the bond and stock markets, although this is by no means a straightforward exercise. I have outlined in my forthcoming paper (fn. 13) one possible way in which this hypothesis might be tested. Regarding the assumption about Fed behavior, the equation that I have estimated to explain Fed behavior seems quite good by conventional statistical standards, although again conventional tests of individual equations must be interpreted with considerable caution.

Tests of the significance of individual coefficient estimates do not allow one to test one model against another (unless the models are nested). I have, however, recently proposed a method for estimating the uncertainty of a forecast from an econometric model that allows one to make comparisons of predictive accuracy across models (see my 1978c paper), and I have applied this method to three other models besides my own (1978d). The three other models are the classical macro-econometric model of Sargent (1976), the six-equation unconstrained vector autoregression model of Chistopher A. Sims, and a "naive" model in which each variable is regressed on a constant, time, and its first eight lagged values. These three models have quite different policy implications from mine, and so estimating the accuracy of my model against these provides one test of the hypothesis that anticipated government actions affect real variables. In Sargent's model anticipated government actions have no effect on real output, and in Sims' model and in the naive model there are no exogenous variables.

Space limitations prevent a detailed discussion of the method of comparison. The

method accounts for the four main sources of uncertainty of a forecast: uncertainty due to 1) the error terms, 2) the coefficient estimates, 3) the exogenous variable forecasts, and 4) the possible misspecification of the model. It is based on successive reestimation and stochastic simulation of the model. Because it accounts for all four sources, it can be used to make comparisons across models.

A sampling of the comparison results is as follows. For real *GNP* the estimated standard errors of the eight-quarter-ahead forecast, taking into account all four sources of uncertainty, are 4.74 percent for the naive model, 5.10 percent for Sargent's model, 7.79 percent for Sims' model, and 2.27 percent for my model. For the eight-quarter-ahead forecast of the *GNP* deflator the corresponding estimated standard errors are 6.20, 8.53, 6.26, and 3.48 percent; and for the eight-quarter-ahead forecast of the unemployment rate the corresponding errors in percentage points are 2.19, 1.88, 2.23, and 0.71. Although these results are quite tentative, they do seem to indicate that my model is more accurate than the other three with respect to these three variables. So as not to leave the impression that I feel I have found the ultimate model, it should also be pointed out that my model is not as accurate as either the naive model or the Sims model with respect to forecasts of the money supply. The estimated standard errors for the eight-quarter-ahead forecast are 3.70 percent for the naive model, 6.79 percent for Sims' model, and 7.50 percent for my model. (The money supply is exogenous in Sargent's model.) Also, my model is not as accurate as the naive model with respect to forecasts of the nominal wage rate. The estimated standard errors for the eight-quarter-ahead forecast are 2.04 percent for the naive model, 5.69 percent for Sims' model, and 4.16 percent for my model. (The wage rate is not a variable in Sargent's model.)

V. Summary and Conclusion

The main points of this paper can be summarized as follows:

1. There is strong theoretical justification from microeconomics for the proposition that even anticipated government actions affect real variables. On a macro-economic level this proposition has received some support in my work in the sense that a number of "micro-economic" explanatory variables are significant in my estimated equations.

2. The possible existence of disequilibrium in the economy provides another justification for the effectiveness of government policies. The proposition that disequilibrium at times exists in the economy has received some support in my work through the significance of my "disequilibrium" variable (*ZJ*).

3. Policy effects in my model are sensitive to assumptions about Fed behavior and about expectations in the bond and stock markets. They are not sensitive to whether or not government debt is treated as wealth by households. It may be possible in the future, as outlined in my forthcoming paper, to test the assumption of rational expectations in the bond and stock markets, but it is unlikely that the macro data can be used to decide how government debt is treated by households. The equation that I have estimated to explain Fed behavior, an equation in which the Fed is estimated to lean against the wind, appears to be good when judged by conventional statistical standards.

4. A method that I have proposed for estimating the expected predictive accuracy of econometric models indicates that my model is more accurate than Sargent's model, Sims' model, and a naive model with respect to forecasts of real *GNP*, the *GNP* deflator, and the unemployment rate. These results thus provide some tentative support for the proposition that even anticipated government actions can affect real variables.

To conclude, it is interesting to speculate what the status of the debate about the effectiveness of government policies will be forty-three years from now in 2021. By this time 172 more quarterly observations will have been generated, and my hope is that the use of these additional data and methods like the one I have proposed for comparing models will have considerably narrowed the range of disagreement. At the least, one would hope that we will have advanced beyond the point where the best model available is only fair.

REFERENCES

R. J. Barro, "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

——— **and H. I. Grossman,** "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 82–93.

Ray C. Fair, *A Model of Macroeconomic Activity, Volume I: The Theoretical Model*, Cambridge, Mass. 1974.

———, *A Model of Macroeconomic Activity, Volume II: The Empirical Model*, Cambridge, Mass. 1976.

———, (1978a) "The Sensitivity of Fiscal-Policy Effects to Assumptions about the Behavior of the Federal Reserve," *Econometrica*, Sept. 1978, *46*, 116s–79s.

———, (1978b) "A Criticism of One Class of Macroeconomic Models with Rational Expectations," *J. Money, Credit, Banking*, Nov. 1978, *10*, 411–17.

———, (1978c) "Estimating the Expected Predictive Accuracy of Econometric Models," Cowles Foundation disc. paper no. 480, Yale Univ., rev. Oct. 1978.

———, (1978d) "An Analysis of the Accuracy of Four Macroeconometric Models," Cowles Foundation disc. paper no. 492, Yale Univ., Aug. 11, 1978.

———, (1978e) "The Fair Model as of April 15, 1978," mimeo., Yale Univ. 1978.

———, "An Analysis of a Macro-Econometric Model with Rational Expectations in the Bond and Stock Markets," *Amer. Econ. Rev.*, forthcoming.

H. I. Grossman, "Why Does Aggregate Employment Fluctuate?," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 64–69.

John Maynard Keynes, *The General Theory of Employment, Interest, and Money*, New York 1939.

R. E. Lucas, Jr., "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

T. J. Sargent, "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment," *Brookings Papers*, Washington 1973, *2*, 429–80.

———, "A Classical Macroeconometric Model for the United States," *J. Polit. Econ.*, Apr. 1976, *84*, 207–37.

——— **and N. Wallace,** "Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

C. A. Sims, "Macro-Economics and Reality," paper presented as the 1977 Fisher-Schultz Lecture, Vienna, Sept. 1977; rev. Nov. 1977.

Jan Tinbergen, *Business Cycles in the United States of America, 1919–1932*, Geneva 1939.

APPLIED WELFARE THEORY

Equilibrium and Welfare in Unregulated Airline Markets

By JOHN C. PANZAR*

Recently, the Civil Aeronautics Board has begun removing the restraints on price competition which have shackled and sheltered *U.S.* airlines for forty years. It becomes important therefore, to develop an understanding of how *unregulated* airline markets can be expected to perform. While the functioning of *regulated* airline markets has received detailed theoretical and empirical analysis (see George Douglas and James Miller, for example), relatively little theoretical effort has been expended on the unregulated case. The prevailing notion seems to be that the industry would perform competitively, since it is generally agreed that airline technology exhibits constant returns to scale.

Unfortunately, this comfortable view fails to take account of the product differentiation effect resulting from variation in flight departure times, and the effects of flight frequency and load factor on service quality. Incorporating these realities results in a monopolistically competitive model of airline markets. In this paper I introduce and analyze such a model. The basic results are that 1) when the *direct* benefits (to consumers) of increasing flight frequency are exhausted, socially optimal choices of price and frequency result in zero profits for the industry, *but* 2) a noncooperative, free entry equilibrium always results in higher prices, lower load factors, and greater frequency than are socially optimal.

I. The Model

The demand for air transportation in a stylized city pair market is assumed to depend upon the ticket price p, and two aspects of service quality: flight frequency n and the load factor L. Consumers' desired departure times are assumed to be uniformly distributed over a circular schedule of duration T. The demand density at any point on this circle is $x(\rho)$, $x' < 0$; where ρ, the "full price" of air travel, is given by

$$(1)\quad \rho = p + h(\text{t}) + g(n, L)\ h' \geq 0,$$
$$\frac{\partial g}{\partial n} = g_n \leq 0,\ \frac{\partial g}{\partial L} = g_L > 0,\ \frac{\partial^2 g}{\partial L^2} > 0$$

where t is number of minutes between desired departure time and that of the most convenient flight.

Equation (1) assumes, as is common in the literature, that consumers are able to place a dollar value on nonprice service attributes. The specification employed here is based upon Douglas and Miller's concepts of frequency delay and stochastic delay. Thus, $h(\text{t})$ represents the value of the inconvenience caused by taking a flight which does not depart at one's desired departure time (frequency delay); and g values the inconvenience resulting when one cannot obtain a seat on the most convenient flight (stochastic delay). While no stochastic structure will be explicitly introduced, it is assumed that one's probability of being denied a seat, and hence stochastic delay costs, will be an increasing and convex function of the average load factor. Since it may not be possible to gain a

*Bell Laboratories. The views expressed are my own and do not necessarily reflect those of Bell Laboratories or the Bell System. I would like to thank R. D. Willig for helpful comments and suggestions. Because of space limitations, only the basic structure of the model and the major results are presented. Additional results and a more detailed formulation and analysis of the model are contained in my unpublished working paper. Alternative approaches to this problem are contained in unpublished papers by George W. Douglas and Gary J. Dorman.

seat on the most convenient flight, the length of the wait until the next flight will also be a component of stochastic delay. The magnitude of this delay clearly decreases with the number of flights. Note that the present specification does not require that the delays in question be valued at a uniform, constant rate.

This specification also permits an important modification of the full price theory of consumer behavior. It allows for the possibility that consumers may be insensitive to small discrepancies between actual and desired departure times. This might result, for example, from the unavoidable uncertainties involved in judging travel time to and from the airport. Figure 1 depicts an h function of this type. Thus, when the interval between flights is sufficiently small (i.e., less than $2\bar{t}$), consumers incur no frequency delay costs. A similar argument explains why it may also be reasonable for g_n to equal zero for large n. Operationally, these suppositions have the testable implication that in dense, heavily scheduled markets, further increases in frequency have no effect upon demand when price and load factor are held constant.

Following the conventional wisdom, the production process is assumed to exhibit constant costs. The firm incurs a cost of c for each passenger carried and a cost $b(k)$ of flying a plane with k seats. Since there are no economies resulting from operating multiple flights the analysis that follows proceeds under the assumption that each firm operates exactly one flight. Because of the underlying symmetry of the model, only the properties of *symmetric* equilibria will be analyzed; that is, flights will be evenly spaced around the circle and charge the same price. This allows one to focus on pricing and entry issues without dealing with the difficult (and interesting) problems of locational rivalry. At any symmetric equilibrium the number of passengers carried by a representative flight is given by

$$(2)\qquad X(p,n,L)= 2\int_0^{T/2n} x[p + h(t) + g(n,L)]\,dt$$

Clearly, the market segment served by each

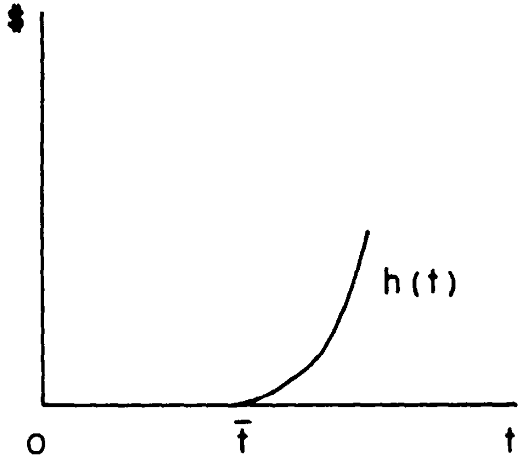


FIGURE 1

flight will, on each side, be one-half the time to the next flight.

II. Welfare Maximization

In order to develop a benchmark by which to evaluate the unregulated market equilibrium, I shall characterize the price and number of flights which would be chosen by a planner attempting to maximize social welfare (W), taken to be the sum over all flights of the consumers' surplus areas under the full price demand curve (S) and profits per flight (π):

$$(3)\qquad W = nS + n\pi = 2n\int_0^{T/2n}\int_{p+h(t)+g(n,L)}^{\infty} x(\rho)\,d\rho\,dt + 2n(p-c)\int_0^{T/2n} x\,dt - nb$$

The load factor L is an important determinant of W which is not under the planner's direct control, being determined, via consumers' preferences, from the relationship between price and available capacity. That is, $L(p, n, k)$ is defined implicitly by (2) and the identity $X = kL$.

Maximizing W with respect to price requires

$$(4)\quad \frac{dW}{dp} = n\{(p - c)X_p + g_L L_p\,[(p - c)X_p - X]\} = 0$$

where subscripts indicate partial differentiation. Using the Implicit Function Theorem, one finds that $L_p = X_p/D$ and $L_n = X_n/D$, where $D = k - g_L X_p$. Therefore, (4) can be rearranged to

$$(5)\quad p^* = c + g_L \frac{X}{k}$$

Equation (5) can be readily interpreted. It requires that the optimal fare p^* be set equal to direct marginal per passenger costs plus the delay costs an additional passenger imposes upon customers already on the flight. (To see this, note that Xg_L is the increase in per flight delay costs resulting from a small increase in load factor, while $1/k$ is the increase in load factor resulting from a small change in X, the number of passengers per plane.) In other words, we reach the usual conclusion that welfare maximization requires the equality of price and marginal social cost.

In order to examine the planner's choice of the number of flights n^*, we differentiate W with respect to n, obtaining

$$(6)\quad \frac{dW}{dn} = \pi + S - \frac{T}{n}s - ng_n X + n\{(p - c)X_n + g_L L_n[(p - c)X_p - X]\} = 0$$

where s is the surplus of consumers whose desired departure times are furthest from the flights they take. In view of (4), and the fact that $L_n = (X_n/X_p)L_p$, the bracketed term in (6) vanishes. Rearrangement then yields

$$(7)\quad \pi(p^*, n^*) = n^* g_n X - \left(S - \frac{T}{n^*}s\right) \leq 0$$

Thus, welfare optimal choices of p and n result in nonpositive profits for the representative firm. While in most monopolistic competition models firm profits are negative at the social optimum, here, they *may* be zero. This will occur when $g_n = 0$ and when the surplus realized by consumers with the "furthest off" desired departure time equals the average surplus of all consumers on a flight. When this happens, each consumer obtains the utility he would receive were his desired departure time actually offered; i.e., $T/2n^* < \bar{t}$. In the standard monopolistic competition model, this could occur only by an infinite number of products being offered. Here, the same result can be achieved with a finite number of flights because the model is rich enough to reflect the intuitively plausible possibility that schedules may be sufficiently dense that further increases in frequency have no *direct* effect on market demand.

If a full welfare optimum can be achieved in conjunction with zero profits for the representative firm, it provides a natural benchmark for evaluating the performance of an unregulated market. The remainder of the paper will address this issue, maintaining the structural assumptions which make possible a zero profit welfare optimum.

III. Nash Equilibrium

The unregulated equilibrium which will be analyzed results from firms (flights) setting their prices noncooperatively to maximize profits, with free entry ensuring that profits are zero in equilibrium. Assuming that the relevant second-order conditions are satisfied, this Nash equilibrium is characterized by

$$(8)\quad (p - c)X^e_p + X^e = 0$$

$$(9)\quad (p - c)X^e - b = 0$$

where $X^e(p)$ is the demand curve *perceived* by the representative firm under the assumption that its rivals' prices remain unchanged.

Since we are examining equilibria under the assumption that a firm has no purely locational advantage over its nearest rivals (i.e., $T/2n < \bar{t}$), it might appear that $X^e(p)$ would be discontinuous, and that a firm could anticipate capturing all the customers of a rival by slightly undercutting its price. Such is not the case, however, because as a firm attracts customers from its rival, its load factor increases relative to that of its rival, allowing full prices to be reequilibrated. The effect of our representative firm's price reduction does not end there, however, since the reduction in the load factor of immediate

rivals makes their flights more attractive relative to those of *their* rivals, resulting in further load factor changes; and so on around the circle. Modelling such an equilibrium system would, in general, be quite complex. However, the task is manageable when one restricts attention to symmetric equilibria in which all firms are equidistant and, initially, charging the same price. In that case, it can be shown that $X_p^e = -k(nD - k)/nDg_L$. Substituting this into (8) yields

$$(10)\quad p^e = c + \frac{g_L X^e}{k}\left[\frac{nD}{nD - k}\right]$$
$$\equiv c + \frac{g_L X^e}{k} \cdot \theta$$

with θ clearly greater than unity.

Equation (10) reveals that, in contrast to the welfare-maximal price, the price resulting from noncooperative rivalry exceeds the marginal social cost of an additional passenger. In order to make a direct comparison between p^* and p^e one must recognize that the relevant load factors L^* and L^e will also differ. However, using (5), (10), and the zero profit condition it is possible to show that $p^e > p^*$, $n^e > n^*$, and $L^e < L^*$. Figure 2 depicts the relative locations of the Nash equilibrium (N) and the welfare optimum (W). Equilibrium in an unregulated airline market results in a higher price, more flights, and, hence, a lower load factor than would prevail at a zero profit welfare optimum.

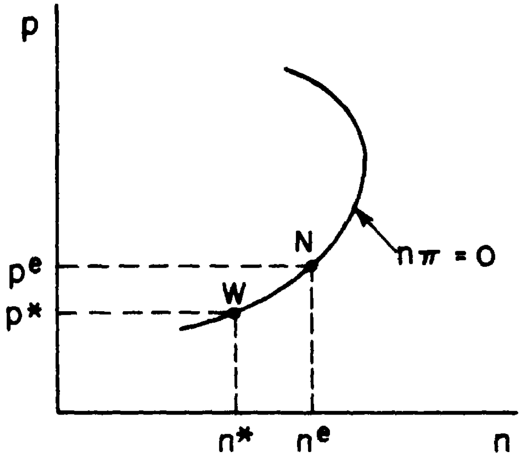


FIGURE 2

from the welfare optimum. It can be demonstrated that the magnitude of this divergence (as measured by θ) depends quite directly on the structural effects of load factor changes on market demand. As noted earlier, whether or not a market admits the possibility of a zero profit welfare optimum depends upon the structural effects of frequency on market demand. Therefore intelligent policy formulation clearly requires accurate estimation of structural air travel demand functions capable of isolating the effects of frequency and load factor as well as price.

IV. Conclusions

The results of the analysis have important implications for both public policy and the direction to be taken in future empirical research. The divergence between the Nash equilibrium and a zero profit welfare optimum suggests that, *in theory*, a policy of price regulation and free entry would yield higher welfare than one of complete deregulation.

However, whether or not one wishes to retain the current unwieldy regulatory apparatus in order to improve market performance must ultimately depend upon the *quantitative* divergence of the unregulated equilibrium

REFERENCES

G. J. Dorman, "Airline Competition: A Theoretical and Empirical Analysis," unpublished doctoral dissertation, Univ. California-Berkeley 1976.

George W. Douglas, "Equilibrium in a Deregulated Air Transport Market," paper presented at Seminar on Problems of Regulation and Public Utilities, Dartmouth College, Aug. 21, 1972.

——— **and James C. Miller**, *Economic Regulation of Domestic Air Transport: Theory and Policy*, Washington 1974.

J. C. Panzar, "A Theory of Unregulated Airline Markets," unpublished paper, Bell Laboratories 1978.

The Economic Gradient Method

By ROBERT D. WILLIG AND ELIZABETH E. BAILEY*

The economic gradient method is designed to provide guidance for the analysis and recommendation of change when information is available only over so narrow a range as to preclude a credible calculation of the global optimum. Using as data the gradients of the relevant functions evaluated at the current point of operation, the procedure calculates the direction of change from the status quo that yields the greatest feasible local rate of increase in the objective function of the decision maker. The calculation should be viewed as a benchmark because proposed movements of practical size must be assessed for feasibility. Under structural assumptions (standard for second-order conditions), the procedure can also be used to obtain an upper bound on the gain from effecting any particular (nonlocal) set of feasible changes in the decision variables.

In Section I, we present the economic gradient method in a general form. We specialize the formulae in Section II to apply to pricing under a budget constraint with aggregate consumer welfare as the objective function.[1] Pilot empirical applications to U.S. Postal Service and long distance telephone rates are summarized in Sections III and IV.

I. The Economic Gradient Method

The economic gradient method requires the specification of a metric to measure the distances between vectors of decision variables, x. Here, we work with the weighted Euclidean metric that measures the distance between x' and x'' as

$$(1) \quad \rho(x', x'') = \left[\sum_i (x_i' - x_i'')^2 w_i\right]^{1/2}, \; w_i > 0$$

The more a given change in x_i contributes to the decay in knowledge about the local behaviors of the objective and constraint functions, $F(x)$ and $G(x)$, the larger should w_i be chosen. Two natural specializations of (1) are the Euclidean metric with $w_i = w_j$, and the percentage change Euclidean metric with $w_i = (1/x_i^o)^2$, where x^o denotes the current, status quo, values of the decision variables. The latter metric is appropriate when equal percentage changes in decision variables cause equal losses of information.

The economic gradient method is based on the following program: maximize $F(x)$ subject to $G(x) \geq G^o$ and $\rho(x, x^o) \leq t$. The solution, $x^*(t)$, is the best feasible set of values of the decision variables within the distance t of the status quo. $(x^*(t) - x^o)/\rho(x^*(t), x^o)$ is a unit vector pointing in the best direction of change within distance t. In the limit, as t approaches 0, this is the best local direction of change from x^o. The gain in the objective function averaged over the permitted distance moved is $[F(x^*(t)) - F(x^o)]/t$. In the limit as t goes to 0, this is $dF(x^*(t))/dt|_{t=0}$, the maximum feasible local rate of increase in the objective function.

Using (1), the Lagrangian for the basic program can be written as $L = F(x) + \lambda(G(x) - G^o) + \gamma[t^2 - \Sigma(x_i - x_i^o)^2 w_i]$. Denoting partial derivatives with respect to x_j by the subscript j, and assuming that F and G are twice differentiable, the Kuhn-Tucker conditions are

$$(2) \quad F_j(x^*) + \lambda^* G_j(x^*) - 2\gamma^* w_j(x_j^* - x_j^o) = 0$$

$$(3) \quad \lambda^*[G(x^*) - G^o] = 0, \quad \lambda^* \geq 0, \; G(x^*) \geq G^o$$

$$(4) \quad \gamma^*\left[t^2 - \sum_i (x_i^* - x_i^o)^2 w_i\right] = 0, \quad \gamma^* \geq 0, \; t^2 \geq \sum_i (x_i^* - x_i^o)^2 w_i$$

As $t \longrightarrow 0$, $x^*(t)$ approaches x^o and

*Professor of economics and public affairs, Princeton University, and member, Civil Aeronautics Board, respectively.

[1] See Robert Dansby and Willig, and Janusz Ordover and Willig for other applications.

$F_j(x^*(t))$, $G_j(x^*(t))$ and λ^* approach $F_j(x^o) \equiv F_j$, $G_j(x^o) \equiv G_j$, and λ, respectively, $\lambda \geq 0$.

Assuming that the first-order conditions for optimality unconstrained by distance moved are not satisfied at x^o, there is no $\lambda \geq 0$ such that $F_j + \lambda G_j = 0$ for all j, and (2) and (4) show that $\gamma^* > 0$ for small t. Then, for such t, (4) yields

$$(5) \qquad t^2 = \sum (x_i^* - x_i^o)^2 w_i$$

Solving (2) algebraically for $x_j^* - x_j^o$, substituting into (5), rearranging, and taking the limit as $t \longrightarrow 0$ gives

$$(6) \quad \lim_{t \to 0} 2\gamma^* t = [\sum (F_j + \lambda G_j)^2 / w_j]^{1/2} \equiv I > 0$$

Rearranging (2) and using (6) yields

$$(7) \quad dx_j^*(t)/dt\,|_{t=0} \equiv \lim_{t \to 0} \left[\frac{x_j^*(t) - x_j^o}{t}\right] = \frac{F_j + \lambda G_j}{I w_j}$$

Note that since $G(x^o) = G^o$ and $G(x^*(t)) \geq G^o$, $dG(x^*(t))/dt\,|_{t=0} \geq 0$. If λ were 0, (7) shows that this would hold iff $\Sigma F_j G_j / w_j \geq 0$. If λ were positive, then, for small t, $\lambda^* > 0$ and $G(x^*(t)) = G^o$, $dG(x^*(t))/dt\,|_{t=0} = 0$, and (7) would yield

$$(8) \qquad \lambda = -\sum (F_j G_j / w_j) / \sum (G_j^2 / w_j)$$

Thus, $\lambda > 0$ iff $\Sigma (F_j G_j / w_j) < 0$, and for brevity, we proceed under this plausible assumption.

The results are most transparently cast in terms of the ratios familiar from first-order conditions for constrained optimality: $R_j = F_j / G_j$. Rearranging (7) and (8), and relabelling λ with $-R$, we have

$$(9) \quad R = \sum_j R_j [(G_j^2 / w_j) / \sum_i (G_i^2 / w_i)]$$

$$(10) \quad dx_j^*(t)/dt\,|_{t=o} = (1/I)(G_j / w_j)[R_j - R^*]$$

Equation (9) defines R^* as a weighted average of the first-order ratios calculated at x^o, R_j. Equation (10) gives the components of the unit vector (in the metric (1)) pointing in the locally best direction of change from x^o.

Application of the envelope theorem to the basic program gives

$$(11) \quad dF(x^*(t))/dt\,|_{t=o} = I = [\sum (G_j^2 / w_j)(R_j - R^*)^2]^{1/2}$$

This form of I is proportional to the standard deviation of the R_j's, relative to the weights G_j^2 / w_j.

Straightforward calculations show that $F(x^*(t))$ is a concave function of t if either (a) $F(x) + \lambda G(x)$ is concave in x at $x^*(t), \lambda^*$ or (b) $F(x)$ is concave and $G(x)$ is quasi concave in x at $x^*(t)$. In these cases, $t[dF(x^*(t))/dt\,|_{t=o}] \geq F(x^*(t)) - F(x^o)$. Since, by definition, $F(x^*(\rho(x, x^o))) \geq F(x)$ for $G(x) \geq G^o$, we have this result:

THEOREM 1: *If* (a) *or* (b) *hold for* $0 \leq t \leq \rho(x, x^o)$, *and if* $G(x) \geq G^o$, *then*

$$(12) \qquad F(x) - F(x^o) \leq I\rho(x, x^o)$$

where I *and* ρ *are defined in* (11) *and* (1).

The inequality in (12) allows an upper bound on the gains from any particular move to be calculated from only current data. If this upper bound fails to exceed exogenous costs of the move, it can be inferred that the move is undesirable.

II. Ramsey Pricing and the Economic Gradient Method

In this section we specialize our results to Ramsey pricing; that is, pricing under a budget, or net revenue constraint, with maximization of social welfare as the objective.[2] Let the constraint be $\Pi(p) \equiv \Sigma p_j y_j - C(y) \geq \Pi^o$, where p is the vector of prices to be set, y is the vector of corresponding quantities demanded at prices p, $C(y)$ is the total cost associated with outputs y, and $\Pi(p)$ is the net revenue, or that portion of profit or producers' surplus relevant to the analysis. We model the social welfare objective function $V(p)$ as having the local behavior of $\Pi(p)$ plus aggregate consumer's surplus. Differentiation with respect to p_j yields

[2]See Frank Ramsey. See also William Baumol and David Bradford. The authors (1977a) describe ways to expand "Ramsey analysis" to take account of externalities, redefinition of the product set, and so on. The study is extended to include income distributional effects in the authors (1977b).

(13) $\Pi_j = y_j(1 - \alpha_j)$, and $V_j = -y_j\alpha_j$, where

$$\alpha_j \equiv \sum_i [(p_i - C_i(y))/p_i][-\frac{\partial y_i}{\partial p_j}\frac{p_j}{y_i}][\frac{p_i y_i}{p_j y_j}]$$

The first-order conditions for feasible prices to be globally optimal for the specified program (Ramsey optimal) are that the *Ramsey numbers*, α, of all of the outputs be between 0 and 1 and be equal to one another. With zero cross elasticities of demand, this is the "inverse elasticity rule" and α_j is the proportional deviation between the jth price and marginal cost multiplied by the own-price elasticity of demand for the jth output. When the Ramsey numbers at the current point of operation are not all equal, optimality has not been reached and social welfare could be increased without diminution of Π by suitable adjustment of prices.

The gradient method supposes that the revenue constraint is satisfied with equality at the status quo and that it remains binding in the locally best direction of price changes.[3] Utilizing the percentage change Euclidean metric, we have

$$(14) \quad (1/p_j^o)dp_j^*/dt|_{t=0} = kp_j^o y_j^o(\alpha^* - \alpha_j)$$

where k is a common positive proportionality factor. Here, α^* is the *critical Ramsey number*. Goods with Ramsey numbers larger than α^* have their prices decreased in the locally best direction and inversely. In this direction, a larger proportional price change is indicated for a good having a relatively large current value of sales ($p_j^o y_j^o$) and having a relatively large deviation ($\alpha^* - \alpha_j$).

The critical Ramsey number is a weighted average of the Ramsey numbers of the goods whose prices are under consideration:

$$(15) \quad \alpha^* = \sum_j \alpha_j((p_j^o y_j^o)^2(1 - \alpha_j)/\sum_i (p_i^o y_i^o)^2(1 - \alpha_i))$$

Each weight in the average increases with the value of the good's sales. The maximum feasible local rate of increase in social welfare is given by

$$(16) \quad I = (1/(1 - \alpha^*)) \cdot [\sum(p_j^o y_j^o)^2(\alpha^* - \alpha_j)^2]^{1/2}$$

It is larger the more diverse and the larger are the Ramsey numbers and the higher are the values of the sales of the goods with α_j's that deviate from α^*.

III. Application to Postal Service Rates

The U.S. Postal Service (*USPS*) is the first enterprise in the United States to publicly make a serious effort to include Ramsey pricing methods in its rate setting. We summarize and use their pilot study,[4] as submitted in the testimony of Bernard Sobin, Docket No. R-74-1.

This study contrasted 1974 *USPS* rates with those that were calculated to be Ramsey optimal for five categories of mail in that year. For each category, the price was taken to be the average revenue per piece, calculated by averaging over mail of different weights. For each class of mail the marginal cost was assumed to be constant and equal to average attributable cost, where attributable costs were defined as those which vary with mail volume on a year-to-year basis. Cross elasticities of demand were assumed to be negligible and the then current own elasticities were in part determined judgmentally. Ramsey optimal prices were calculated using a specification of linear demand curves, and with the constraint that net revenue (revenue less "attributable cost" of the five categories) be held constant.

Table 1 displays the parameters and results of the *USPS* Ramsey study. The percentage change Euclidean distance between the *USPS* and Ramsey prices is .29, $\alpha^* = .055$, and $I =$ \$55.8 million. The move to the calculated optimal rates required a 5.6 percent increase for first class mail and decreases from 8 to 17 percent for the other classes. Although the calculations show substantial changes in

[3]It suffices that at the current point an increase in any price would increase Π and decrease V.

[4]Subsequent work by the *USPS* has greatly increased in sophistication.

TABLE 1—THE SHIFT FROM *USPS* RATES

	USPS					Calculated Ramsey Optimal		Benchmark Gradient Method
	Price (¢ per piece)	Quantity (Millions of Pieces)	Own-Price Elasticity	Marginal Cost (¢ per piece)	Ramsey Number	Price Changes (¢ per piece)	Change in Consumer's Surplus (Millions of $)	Price Change (¢ per piece)
First Class	11.27	52,715.3	.100	5.25	.053	+.63	−332.6	+.68
Second Class	7.37	4,264.7	.450	5.76	.098	−.61	+26.7	−.56
Third Class	6.92	16,248.0	.175	3 30	.092	−1.20	+197.7	−1.60
Parcel Post	137.97	427.2	.250	83.66	.098	−20.32	+88.4	−19.44
Special Rate Fourth Class	63.36	275.5	.276	39.38	.105	−9.76	+27.5	−3.07

consumer's surplus for each class of mail, the calculated net gain in consumer's surplus from the move was only $7.7 milllion.

In the postal service study the use of approximations was believable, since the current point was "close" to the calculated optimum. Thus, this study provides us with a natural test of the reasonableness of the gradient method. We normalized our benchmark calculation to the same distance, 0.29, as appeared in the *USPS* study. The price changes displayed in Table 1 are similar for all but fourth class to the changes from the *USPS* to the calculated Ramsey prices. The latter difference is explained by the low weight given to low-value services in the percentage change metric. The upper bound given by (12) on the welfare gain from price changes of this distance is $.29 \cdot I = \$16.2$ million $> \$7.7$ million.

In the next application, we find ourselves in a situation where approximations are not believable and so we must rely directly on the economic gradient method.

IV. Application to Long Distance Telephone Rates

This pilot study analyzes the 1973 prices of the direct distance dial (*DDD*) interstate message toll telephone services. The services consist of point-to-point calls within any of twenty-one mileage bands (short haul to long haul) and during any of the three rate periods labelled day (just weekdays), evening, and night (these rates also held during weekends and holidays). For each of these sixty-three service categories, price was taken to be the average revenue per call. Marginal costs were assumed to be constant over the ranges studied. We assume zero cross elasticities of demand between each of the services analyzed and all other telecommunications services.[5]

Figure 1 displays experimental data on the 1973 proportional deviation between price and marginal cost and on the own-price elasticity of demand, the components of the Ramsey numbers, for the sixty-three services. Separate curves connect the points representing day, evening, and night calls. While the elasticities and proportional markups were inversely related across times of day, they both varied directly with length of haul. Thus, it is clear that the Ramsey numbers increased with the mileage, and that, given the assumptions, there existed price changes in 1973 that would have increased consumer welfare without decreasing the net revenue from these services.

Although we lacked global information on the demand functions, we proceeded to calculate Ramsey optimal prices under the specifi-

[5]See the authors (1977a) for a list of other assumptions and serious caveats.

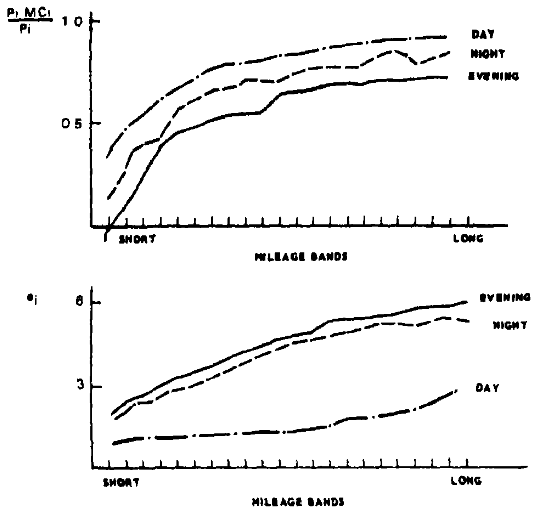


FIGURE 1. ROUGH EXPERIMENTAL DATA

cation of linear and of constant elasticity demands. Unfortunately, the calculated optima involved price changes that were so large (over 140 and 400 percent, respectively) that they eliminated any credibility of our functional assumptions over the range of indicated

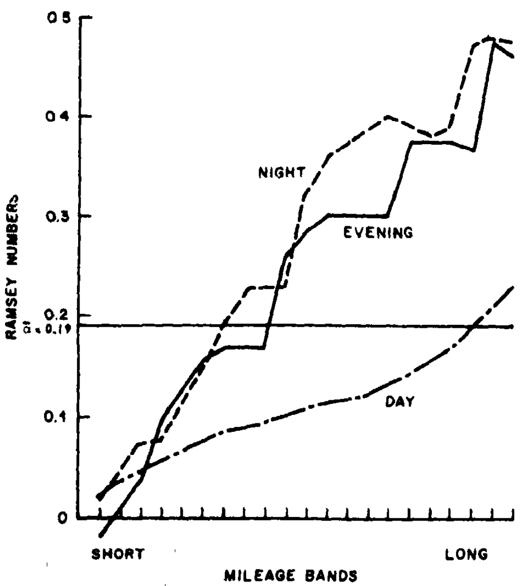


FIGURE 2. RAMSEY NUMBERS OF 1973 *DDD* RATE SCHEDULE

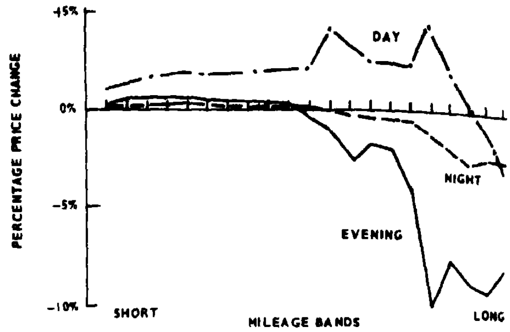


FIGURE 3. BEST DIRECTION OF PRICE CHANGE WITH GREATEST PRICE CHANGE EQUAL TO 10 PERCENT

price changes. Consequently, we applied the economic gradient method, using the percentage change Euclidean metric, as described in Section II. Figure 2 displays the 1973 Ramsey numbers of the services together with the dividing line at the critical Ramsey number of $\alpha^* = .19$. Figure 3 displays the relative price changes in the locally best direction, normalized so that the largest percentage price change is 10 percent.[6]

The upper bound given by (12) and (16) on consumer's surplus gains from feasible price adjustments is \$840,000s for price movements within a distance large enough to hold a change of s percent in each price. For example, a rather large set of feasible price movements might be equivalent in distance to 50 percent changes in each price. Given (12), such a change could not increase consumers' surplus by more than \$42 million. While this figure is less than 1 percent of the 1973 dollar sales of the services analyzed, it should properly be compared to the perceived costs of the price adjustments.

The economic gradient method cannot provide automatic answers to the questions of whether price adjustments are worthwhile and, if they are, exactly what changes should be effected. However, our thesis is that the method is a tool that can reliably be used to better inform the judgments of a decision maker.

[6]Compare Figures 2 and 3 with their counterparts in the authors (1977a)

REFERENCES

W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

R. E. Dansby and R. D. Willig, "Industry Performance Gradient Indexes," *Amer. Econ. Rev.*, forthcoming.

J. A. Ordover and R. D. Willig, "On the Optimal Provision of Journals qua Sometimes Shared Goods," *Amer. Econ. Rev.*, June 1978, *68*, 324–38.

F. P. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, *37*, 47–61.

B. Sobin, *Rebuttal Testimony*, Docket No. R-74-1, "... USPS Changes in Rates of Postage and Fees for Postal Services," Washington, Aug. 20, 1974.

R. D. Willig and E. E. Bailey, (1977a) "Ramsey Optimal Pricing of Long Distance Telephone Services," in John T. Wenders, ed., *Pricing in Regulated Industries: Theory and Application*, Denver 1977.

——— and ———, (1977b) "Income Distributional Concerns in Regulatory Policy Making," in *Proceedings of NBER Conference on Public Regulation*, forthcoming.

Quasi-Walrasian Theories of Unemployment

By GUILLERMO CALVO*

In this paper I will discuss recent attempts to explain labor unemployment on the basis of models that will, somewhat imprecisely, be termed "quasi-Walrasian." These are models where agents perceive themselves as infinitesimal micro units, unable to exercise any kind of "market power"; the agents set the variables under their control in order to maximize their objective functions (utility, profit, etc.). In contrast to the so-called "disequilibrium" or "fix-price" models (see Allan Drazen) there is no exogenous rationing or price inflexibilities. However, they differ from standard Walrasian models (for example, see Kenneth Arrow and Frank Hahn, chs. 1–5) in that agents are not necessarily price takers and, typically, markets are incomplete.

Let us classify the quasi-Walrasian models into two categories: (i) models that rely on costly labor mobility (Section I), and (ii) those that emphasize imperfect information about workers' characteristics, and institutional or legal constraints on labor contracts (Section II). In Section III the possible effects of monetary policy in the context of these types of models will be briefly discussed.

I. Costly Labor Mobility

This approach is presented in the pioneering works of Costas Azariadis (1975) and Martin Baily (1974, 1977). Here we follow Azariadis and consider a situation where workers have to decide *at the beginning* of each period (*ex ante*) the firm to which their labor services will be available *during* the period (*ex post*); *ex post* workers are unable to move to another firm; if they are employed they get the ruling wage at the firm where they positioned themselves; otherwise, they become unemployed with income equal to h (unemployment compensation, self-employment wage, etc.).

Let s, $s = 1, 2, \ldots, S$, be the set of possible states of nature and let $q(s)$ be the probability that everyone assigns to state s *ex ante* (the realization of s is observed *ex post*). Firms are envisioned as competing *ex ante* by offering wage and employment contracts, so that in equilibrium an individual worker's *expected* utility of becoming a member of the "pool" of workers of a given firm is the same as that in any other firm.

The profit for the representative firm in state s, $\pi(s)$ is

$$(1)\qquad \pi(s) = R(n(s), s) - w(s)n(s)$$

where R is the nominal revenue function, and $w(s)$ and $n(s)$ are the nominal wage and employment in state s, respectively.

Assuming the wage-goods price level is constant and equal to one, the *expected utility if s occurs*, $U(s)$, is given by

$$(2)\qquad U(s) = u(w(s), l_w)p(s) + u(h, l_u)\,[1 - p(s)]$$

where u is a von Neumann-Morgenstern utility index, $l_w < l_u$ are the leisure contents associated with working and being unemployed, respectively, and

$$(3)\qquad \text{probability of being employed in state } s \equiv p(s) = \min\{n(s)/m, 1\}$$

m is the size of the labor pool in the representative firm. This asserts that every worker in the pool has the same probability of being employed.

*Associate professor of economics, Columbia University. I wish to thank Costas Azariadis, Allan Drazen, Steven C. Salop, John B. Taylor, Andrew Weiss, and Stanislaw Wellisz for their comments on a previous draft. However, I retain responsibility for opinions and errors. Financial support by the National Science Foundation is also gratefully acknowledged.

Firms are assumed to maximize expected profit by offering a "contract" $\{n(s), w(s)$ $s = 1, 2, \ldots, S\}$ taking into account the fact that in equilibrium the size of the pool will be determined such that

$$(4)\quad \begin{array}{l}\text{expected utility}\\ \text{of belonging to}\\ \text{the firm's pool}\end{array} = \sum_{s=1}^{S} U(s)q(s) = k = \begin{array}{l}\text{expected utility}\\ \text{of belonging to}\\ \text{the pool of any}\\ \text{other firm}\end{array}$$

$$(5)\quad n(s) \leq m$$

(employment cannot exceed the size of the pool)

Assuming that workers are risk averters one can show that optimal $w(s)$ = a constant for all s and, more importantly, that there are configurations under which optimal $n(s) < m$ for some s (see Azariadis, 1975). So the model is capable of explaining the emergence of unemployment and wage rigidity as a *consequence* of the equilibrium outcome of the competitive interaction of the firms and workers who face the reality of costly labor mobility. Furthermore, although the constancy of w depends to a large extent on the specific formulation of the model (in particular on the assumption that firms are expected-profit maximizers), the possibility of unemployment in some states of nature is maintained under a widely different set of assumptions.

Optimal contracts are designed to maximize expected profits and, therefore, only in very special cases will they be consistent with *ex post* profit maximization (even when the firm pays the contractual w). In fact, it can be shown (see the author, 1978a) that the optimum contract requires the firm to behave *ex post* as if its objective were to maximize expected utility (as defined by (2)) subject to a profit target. Thus, if verification of the actual state of nature or enforcement is less than perfect, the firm might try to move towards the *ex post* optimum by either (a) claiming that the state of nature is other than the one that actually happened or (b) simply dishonoring the contract.

Contracts where a firm would not be induced to misrepresent the state of nature were studied by myself and Edmund Phelps. In that paper a contract specifies w as a function of n—named "employment contingent" contracts—and the firm is assumed to *ex post* maximize profits subject to the contracted wage-employment schedule. It was shown that in an optimal contract w is a nondecreasing function of n, and that in certain cases w would in fact be rising with n; furthermore, it was also argued that there are cases where $n(s) < m$ for some s, thus showing that the Azariadis-Baily explanation of unemployment is not necessarily dependent on the firm not behaving as an *ex post* profit maximizer (nor, as a matter of fact, on risk aversion since the example developed by the author and Phelps assumes both firms and workers are risk neutral). Wage contingent contracts will normally result in lower expected profit than those that are indexed to the state of nature and, consequently, firms will try to index their contracts to other easily observable variables that could serve as proxies for some important aspects of the state of nature in order to get closer to the Baily-Azariadis first best. Phelps and I give some examples where firms would find it profitable to make the *real* wage contingent upon the price level, thus suggesting the possible *existence* of a subtle channel by which monetary policy could affect real variables by providing observable information on the state of nature.

When enforceability of contracts is costly, dishonoring contracts becomes a serious problem and raises fundamental doubts on the relevance of contract theory for nonunionized sectors where many aspects of contracts would be "implicit" and hence nonenforceable. I think this is an issue of crucial importance for contract theory because if contracts were not enforceable, what would prevent a firm from *ex post* offering a wage which would make workers just a little better than being unemployed, a policy that would either eliminate unemployment or make its incidence negligible from a practical point of view? Of course, I am not suggesting that "survival" wages are the natural outcome of the Azariadis-Baily set-up, but instead that enforceability (and the related issue of how

explicit labor contracts are) is an aspect of contract theory that needs to be studied more carefully and that, in all probability, will give us further insight on the functioning of the labor (and other) market(s) (see Michael Wachter and Oliver Williamson).

II. Imperfect Information

Theories in this classification assume auction labor markets, but firms find it profitable to offer higher than market-clearing wages because the wage affects the (unobservable or unpunishable) worker's behavior or its quality. Let us study two models that although formally very similar to each other, involve quite different yet complementary micro stories. The first is related to the work of Phelps, Dale T. Mortensen, Steven Salop (1973, 1979) and Joseph Stiglitz.

The first basic assumption is that the quit rate q is such that (for $v \geq \bar{v}$)

$$(6) \quad q = Q(v - \bar{v}),\ Q' < 0,\ Q'' > 0,\ Q(\infty) \geq 0$$

where v is the *real* wage offered by the firm and $\bar{v}$ is the expected real wage if the worker quits or gets fired. (The micro foundation of this assumption appears to be self-evident but has not yet been integrated into a general equilibrium framework.)[1] The second basic assumption is that hiring or training new workers is costly. Assuming, for simplicity, that the marginal labor productivity is a positive constant α, profit at a steady state, π, is given by

$$(7) \quad \pi = [\alpha - v - Q(v - \bar{v})T]L$$

where T is the hiring/training cost per new laborer and L is the total number of laborers. (Also I have set the price of output equal to 1 and assumed that the relevant deflator for calculating v is the price of the firm's output; this will cause no problem in what follows because all firms are assumed identical.) Following Salop (1979) I short circuit the whole dynamic problem that the firm has to face by assuming v and L are chosen so as to maximize π (this would happen at the optimum steady state if the rate of discount were equal to zero). The first-order condition with respect to v yields[2]

$$(8) \quad 1/T = -Q'(v - \bar{v})$$

Thus equation (8) determines the equilibrium "wage differential" $v - \bar{v}$ which we denote z^q. Market equilibrium also requires

$$(9) \quad \pi = 0 \text{ (the zero profit condition)}$$

In order to close the model in a simple fashion we assume

$$(10) \quad \bar{v} = v^q L^q$$

where v^q and L^q are the equilibrium levels of v and "rate of employment" (i.e., employment per unit of active population).[3] Without loss of generality, let us set the number of firms and total active population (exogenous to the model) equal to 1. Thus, in equilibrium $v^q = v$ and $L^q = L$, and hence

$$(11) \quad z^q = v - \bar{v} = v^q(1 - L^q)$$

Now, combining (9) and (11) we get

$$(12) \quad \frac{z^q}{v^q} = \frac{z^q}{\alpha - Q(z^q)T} = 1 - L^q = \text{equilibrium rate of unemployment}$$

It is easy to see that functions could be specified such that $L^q < 1$, i.e., where equilibrium requires a positive rate of unemployment.

At a steady state this model explains only quits, not layoffs; the basic reason for unemployment is the fact that quits can be influenced by v. In their attempt to minimize hiring/training costs, firms drive v above the full-employment level; there is an end to the wage progression because in the process more and more workers become unemployed which creates a downward pressure on $\bar{v}$. Notice that this phenomenon, and hence unemployment, would not arise if the firm could write

[1]The special form is assumed for expositional convenience but it is not essential for the main argument.

[2]Here and in what follows we assume the existence of interior solutions.

[3]More in general $\bar{v}$ will be some function of v^q, L^q, the rates of quits and layoffs, etc., but the central implication of the model would not be affected.

contracts where quitters are punished (legal constraint) or it could identify the quitters from the start and pay differential salaries (information plus legal constraints).

A second type of model is related to my 1977, 1978b papers and my papers with Stanislaw Wellisz (1978a,b). The basic notion here is that supervision is costly (informational constraints) and that punishment for substandard performance has an upper bound (a firm can suspend or fire a worker, but typically imprisonment or cash fines are not feasible (legal constraints)). The author and Wellisz (1978a) show that given a less than perfect supervision level, effort x, will in general be sensitive to changes in v because, for example, an increase in v (given $\bar{v}$) will increase the cost to the worker of being suspended or fired. A similar conclusion on *average* effort could be derived from the author (1977) if workers are heterogeneous in their propensity to "shirk." Following the author (1978b) we assume (for $v \geq \bar{v}$)

$$x = f(v - \bar{v}),\ f(0) \geq 0,\quad f' > 0,\ f'' < 0 \tag{13}$$

so, assuming the same production conditions as above, profit π is defined as follows:

$$\pi = [\alpha x - v]L \tag{14}$$

The first-order condition with respect to v now is

$$\alpha f'(v - \bar{v}) = 1 \tag{15}$$

which determines the equilibrium $v - \bar{v}$, denoted z^s. Once again, equilibrium requires

$$\pi = 0 \text{ (zero profit condition)} \tag{16}$$

Thus, adopting the same normalizations as in the previous model and defining $\bar{v} = L^s v^s$, where L^s and v^s are the equilibrium values of L and v, we get, by (13)–(16),

$$z^s \frac{f'(z^s)}{f(z^s)} = 1 - L^s = \text{equilibrium rate of unemployment} \tag{17}$$

where L^s is the equilibrium rate of employment for this model. Since, by (15), $f'(z^s) > 0$, it follows that in (an interior) equilibrium $L^s < 1$, i.e., unemployment is positive. Notice that this equilibrium is consistent with layoffs (but no quits) because only in special cases will equilibrium wages induce workers to perform at levels at which they would run no risk of being fired (this argument becomes more compelling in a situation where workers are heterogeneous but the firm cannot separate the "hard workers" from the others when it hires them—informational constraints). Notice also, that again in this model it is the behavior of the *employed* workers that keeps the wage rate from falling to its full employment level.

This type of model has also been employed to cast further light on hierarchical structures and labor quality (see the author and Wellisz (1978b)) and seems to point out the possibility of building up a theory of the incidence of cyclical fluctuations on unemployment of different types of labor.[4]

At this point my major reservation with respect to these models is that although they have taught us some of the possible consequences of realistic informational gaps, they do so within a framework where firms are constrained to offer a wage instead of "richer" or more imaginative incentive packages that are common practice in many firms (like seniority clauses, bonuses, etc.). In my 1977 paper, Section 2, there is some discussion on the possibility of a wage cum supervision system (the one that is implicit in the formulation of the last model) being more profitable than one consisting of teams cum incentive wage schedules, but much more research is required.

III. Effects of Monetary Policy

A realistic completion of the above models would take into account nonhuman wealth held by workers. Thus changes in the latter will normally affect risk preference, work-leisure tradeoffs, etc., and, therefore, also the

[4]Andrew Weiss has shown that the sorting effect of wages can also generate an unemployment equilibrium. He shows that in a situation where the expected labor endowment of a hiree is an increasing function of the firm's wage offer, firms may choose not to lower wages when confronted with a queue of job applicants.

equilibrium levels of unemployment. Consequently, a more inflationary policy, for example, will in general affect unemployment through the impact of the former on real wealth (and, in particular, on real cash balances) via Mundell-Tobin-Sidrauski-type of mechanisms.

Another channel would, in principle, exist if unemployment compensation were not perfectly indexed to the price level. In the extreme case where the former is set constant in nominal terms, for instance, the value of h in Section I and the functions for $\bar{v}$ in Section II and, hence, also the equilibrium level of unemployment will be affected by monetary policy.

Finally, Azariadis (1977) has shown in the context of a general equilibrium overlapping-generations model that the variance of money supply is bound to affect expected unemployment if individuals cannot distinguish monetary from real shocks; also in his 1978 paper, he has shown that a similar result holds in a situation where contracts incorporate (variable and endogenous) information lags due to the fact that information is costly to acquire.

REFERENCES

Kenneth J. Arrow and Frank H. Hahn, *General Competitive Analysis*, San Francisco 1971

C. Azariadis, "Implicit Contracts and Underemployment Equilibria," *J. Polit. Econ.*, Dec. 1975, *83*, 1183–202.

———, "A Re-Examination of Natural Rate Theory," mimeo., Univ. Pennsylvania, Nov. 1977.

———, "Naive Contracts and Information Lags in the Monetary Mechanism," paper presented at the 1978 Summer Meetings of the Econometric Society, Colorado.

M. N. Baily, "Wages and Unemployment under Uncertain Demand," *Rev. Econ. Stud.*, Jan. 1974, *41*, 37–50.

———, "On the Theory of Layoffs and Unemployment," *Econometrica*, July 1977, *45*, 1043–64.

G. A. Calvo, "Supervision, and Utility and Wage Differentials across Firms," mimeo., Columbia Univ., July 1977.

———, (1978a) "The Ex-Post Behavior of Firms Offering Optimal Employment Contracts," mimeo., Columbia Univ., Apr. 1978.

———, (1978b) "Involuntary Unemployment and Excess Capacity: an Exploratory Model of Equilibrium and Pure Competition," paper presented at the 1978 Summer Meetings of the Econometric Society, Colorado.

——— **and E. S. Phelps,** "Appendix: Employment Contingent Wage Contracts," in Karl Brunner and Allan H. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conference Series on Public Policy, *J. Monetary Econ.*, Suppl. 1977, 160–68.

——— **and S. Wellisz,** (1978a) "Supervision, Loss of Control, and the Optimum Size of the Firm," *J. Polit. Econ.*, Oct. 1978, *86*, 943–52.

——— **and** ———, (1978b) "Hierarchy, Ability, and Income Distribution," mimeo., Inst. Int. Econ. Stud., Stockholm, June 1978.

A. Drazen, "Recent Developments in Macroeconomic Disequilibrium Theory," paper presented at the 1978 Summer Meetings of the Econometric Society, Colorado.

D. T. Mortensen, "A Theory of Wage and Employment Dynamics" in Edmund Phelps ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970, 167–211.

E. S. Phelps, "Money Wage Dynamics and Labor Market Equilibrium" in his *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970, 126–66.

S. C. Salop, "Wage Differentials in a Dynamic Theory of the Firm," *J. Econ. Theory*, Aug. 1973, *6*, 321–44.

———, "A Model of the Natural Rate of Unemployment," *Amer. Econ. Rev.*, Mar. 1979, *69*, 117–25.

J. E. Stiglitz, "Wage Determination and Unemployment in L.D.C.'s," *Quart. J. Econ.*, May 1974, *88*, 194–227.

M. L. Wachter and O. E. Williamson, "Obligational Markets and the Mechanics of Inflation," disc. paper no. 7, Center Study Org. Innovation, Univ. Pennsylvania, Nov. 1977.

A. Weiss, "An Adverse Selection Model of Layoffs and Excess Demand for Industrial Employment," mimeo., Bell Laboratories 1978.

Staggered Wage Setting in a Macro Model

By JOHN B. TAYLOR*

Few economists now question the validity of the Friedman-Phelps accelerationist hypothesis that the Phillips curve is vertical in the long run—at least as a first-order approximation. Indeed, the once controversial hypothesis is now embodied in basic textbook macro models (see Rudiger Dornbusch and Stanley Fischer, and Robert J. Gordon, for example). This new accelerationist consensus, however, has done little to settle the ongoing debate over aggregate demand policy, where the crucial issues appear to depend on the *short-run* Phillips curve and its dynamic properties. The accelerationist theory provided an elegant and concise representation of the inflationary process for the long run. However, it has proved distressingly unspecific as a framework for the development of short-run dynamics.

Two sources of this incomplete specification have stimulated extensive research in recent years. The first—about which little will be said here—is that the accelerationist theory was not specific about the process of expectation formation. According to the theory, the expected inflation rate π^* should be added to the right-hand side of the Phillips equation. Hence the expectation process determining π^* matters greatly for short-run dynamics. For example, if expectations are formed rationally, then as Thomas Sargent and Neil Wallace have shown (using the Robert Lucas supply model), the Phillips curve will be vertical in the short run as well as the long run from the point of view of aggregate demand policy. On the other hand, if expectations are adaptive—either by assumption or by derivation from a learning model—then the short-run slope might be very flat. But, if this were the only source of ambiguity in the accelerationist model, then it is likely that the controversy over the short-run properties would have been settled quickly: the attractiveness of rational expectations—again as a first-order approximation—has become increasingly evident in theoretical and empirical work.

The second source of imprecision is more troublesome and is unlikely to be resolved quickly. It involves the micro-economic details of wage and price adjustment which are just as much a part of the famous macro "expectations" adjustment, as the expectation formation mechanism itself. While an extremely literal reading of the accelerationist theories would interpret π^* as a pure forecast of inflation independent of the dynamics of wage and price contracts, a more practical reading would suggest that π^* represents the persistence of inflation due to the gradual adjustments of outstanding wage and price contracts to new economic information. Some modelling of this phenomenon can be found in Edmund Phelps (1970), especially his Appendix 1, and in Arthur Okun's contract-based inflation model with accelerationist implications. Empirical work on price and wage equations by Philip Cagan and Michael Wachter has emphasized the dynamic implications of both contracts and expectations. Policy-oriented studies by William Fellner (1978) and George Perry have also taken this view of the accelerationist theory, though with widely differing policy suggestions.

The impact of aggregate demand on inflation and employment is crucially dependent on whether the contract mechanism or the expectation mechanism dominates the persistence effects commonly represented by π^*. Hence, a resolution of the current macroeconomic controversy requires some explicit models to disentangle the two mechanisms theoretically, if not empirically, and to determine how contract length and adjustment

*Columbia University. This research is being supported by the National Science Foundation. This paper was completed while I was a consultant to the Federal Reserve Bank of Philadelphia, which does not necessarily endorse the views expressed. I wish to thank Martin Baily, Philip Cagan, Guillermo Calvo, and Edmund Phelps for useful comments.

speeds affect aggregate demand.

The purpose of this paper is to discuss one such model which focuses on contracts and staggered wage setting with rational expectations. The model is based on some of my recent research (see my 1978, 1979 papers) but is generalized here to permit alternative mixes of expectation and contract effects in the wage equations.

I. Staggered Wage Setting

A property of wage and price contracts which has not typically been emphasized in micro-economic analyses, but which is important from the viewpoint of macroeconomics is that contract decisions are staggered: all contract decisions in the economy are not made at the same point in time. While some months are more popular than others for adjusting wage contracts, these adjustment decisions are generally staggered throughout the year. This property of contract formation is the central feature in the model discussed below.

To make things simple suppose that *wage* contracts last one year and that decision dates are evenly staggered: half the contracts are set in January and half in July. If we let six-month (semiannual) intervals be the period of measurement, and x_t be the *log* of the contract wage for periods t and t + 1, set at the start of period t, then a simple model of contract wage determination is given by

$$(1)\quad x_t = bx_{t-1} + d\hat{x}_{t+1} + \gamma(b\hat{y}_t + d\hat{y}_{t+1}) + \epsilon_t$$

where y_t is a measure of excess demand in period t, ϵ_t is a random shock, and b, d, and γ are positive parameters. The "hat" over a variable represents its conditional expectation based on period t − 1 information. Equation (1) states the assumption that the contract wage set at the start of each semiannual period depends on three factors: the contract wage set in the previous period, the contract wage expected to be set in the next period, and a weighted average of excess demand expected during the next two periods. Since, by assumption, x_t will prevail for two periods, firms and/or unions contemplating a wage adjustment in period t will be concerned with wage rates which will be in effect during periods t and t + 1. Hence both x_{t-1} and $\hat{x}_{t+1}$ are included in the equation. Note that contracts set before period t − 1 and after period t + 1 are not included in the equation. Such contracts do not overlap with the current contract and are therefore not part of the relative wage structure.

The b and d coefficients in equation (1) represent the elasticity of the current contract wage with respect to the previous contract wage and the next contract wage, respectively. Let us assume that $b + d = 1$ so that the current contract decision is homogeneous of degree 1 in these lag and lead contracts. If $b = d = 1/2$ then the lag and lead distribution is symmetric. This has been the parametric assumption used in my previous work and reflects the plausible assumption that current negotiations weight other contracts according to the number of periods that they overlap with the current contract. In this sense, when b and d are equal to 1/2, contract decisions are unbiased. Wage setters look forward to the same degree they look backward. However, I will allow for the possibility of biased weights in this paper by permitting b and d to differ from 1/2. This permits a spectrum of contract determination hypotheses between the extremes of pure backward looking ($b = 1$), and pure forward looking ($d = 1$). As will be demonstrated below the size of b vs. d is important for the dynamic behavior of contracts, and for the sensitivity of wage behavior to excess demand. This importance of forward looking vs. backward looking has been emphasized in a recent paper by Perry in analyzing an hypothesis set forth by Fellner (1976).

In order to derive a dynamic representation for the behavior of the contract wage from equation (1), it is necessary to solve for $\hat{y}_t$, $\hat{y}_{t+1}$, and $\hat{x}_{t+1}$. This involves specifying an aggregate demand relationship and a policy rule. Assume that the excess demand variable y_t is the percentage output gap (that is, the deviation of the *log* of real output from trend), and that the demand for money is given by $m_t = y_t + w_t - v_t$ where the

variables m_t, w_t, and v_t are the *logs* of the aggregate wage level, the money supply and a shock, all measured as deviations from trend. Note that this money demand equation is simply the quantity equation with the wage substituted for the price level. This approximation saves one equation and can easily be modified. If the policy rule for the money supply is the *log*-linear form $m_t = gw_t$, then we can derive the simple aggregate demand relation

$$(2) \qquad y_t = -\beta w_t + v_t$$

where $\beta = 1 - g$. Note that β is a policy parameter indicating the degree of accommodation of aggregate demand to wage changes. The model is closed by noting that w_t is an aggregate of the contract wages x_t and x_{t-1} outstanding at time t. If we use the geometric average, then

$$(3) \qquad w_t = .5(x_t + x_{t-1})$$

By substituting equations (3) and (2) into (1) and taking expectations conditional on t − 1 information we have that

$$(4) \qquad b\hat{x}_{t-1} - c\hat{x}_t + d\hat{x}_{t+1} = 0$$

where $c = (1 + .5\gamma\beta)/(1 - .5\gamma\beta)$. Assuming that x_t is stable yields a solution for x_t of the form

$$(5) \qquad x_t = \alpha x_{t-1} + \epsilon_t$$

$$\text{where} \quad \alpha = \frac{c - [c^2 - 4d(1-d)]^{1/2}}{2d}$$

An equation for the average wage w_t can readily be derived from (5) using (3) and is given by

$$(6) \qquad w_t = \alpha w_{t-1} + .5(\epsilon_t + \epsilon_{t-1})$$

Equations (6) and (2) can be used to address a number of the issues raised above. From the parameter α we can determine how the wage dynamics depends on aggregate demand policy (β), on the sensitivity of wage change to excess demand (γ), and on the degree of forward looking (d).

Note, however, that in this model we cannot identify the two parameters γ and d from a time-series on w_t and y_t without further assumptions. Given such time-series we could easily estimate β and α from equations (2) and (6). However, from the definition of α, these estimates would not determine d and γ uniquely. Of course this identification problem could be surmounted by making additional assumptions or by looking for shifts in policy. For example, additional identifying constraints arise when contracts last for more than two periods. Nevertheless, this potential identification problem should be kept in mind when attempting to estimate the degree of forward looking using aggregate time-series data.

II. Forward Looking Contracts and Aggregate Wage Dynamics

The parameter α in equation (5) characterizes the degree of persistence in aggregate wage behavior. Clearly the persistence will depend on how accommodative aggregate demand policy is to wage contract adjustments which are "too inflationary." This dependence is captured in the model by the relationship between α and β. The higher is β (i.e., the less accommodative is policy) the lower is α (i.e., the less persistent are wage fluctuations). Hence by choosing β large enough, policy can achieve high degrees of stability in the path of aggregate wages. However, since higher values of β result in larger fluctuations in the output gap (see equation (2)), this wage stability must be traded off against real output and employment stability. This stability tradeoff defines the inflation unemployment dilemma in this model.

In order to distinguish between the impact of contract effects and expectations effects on this tradeoff, the parameter d can be varied over its range between 0 and 1. Recall that the lower is d the more backward looking is contract determination and the less important are expectations. For certain values of β and γ, Figure 1 illustrates how the wage dynamics depend on d. As one would expect, smaller values of d are associated with larger values of α. That is, more backward-looking wage determination increases the persistence or the inertia of aggregate wages. The shape of this negative relationship shows that increasing

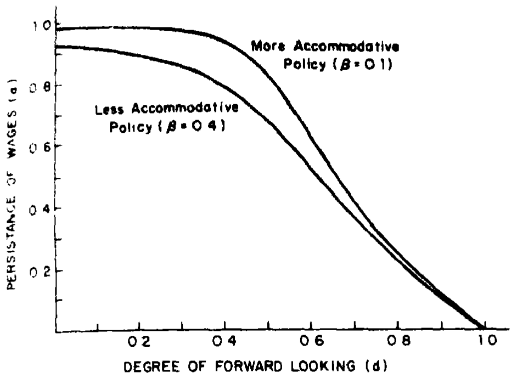


FIGURE 1. THE EFFECT OF FOWARD LOOKING CONTRACTS ON THE PERSISTENCE OF WAGES ($\gamma = 2$)

forward looking (d) from 40–50 percent would reduce persistence substantially. Increasing d from 10–20 percent would only reduce persistence slightly, however.

It can also be shown that the wage-output stability tradeoff depends on d. Because forward looking increases the demand effects on wages, higher values of d improve the tradeoff. This corresponds to the intuitive notion that more forward-looking contract determination increases the impact of aggregate demand policy on wages. Hence, inflation-stabilizing fluctuations in aggregate demand can be smaller and need not last as long.

III. Contract Length, Empirical Regularities, and Micro Foundations

The wage and output dynamics generated by this model share a number of features with the actual behavior of these series and this lends some support to the idea that contract formation as well as expectations is an important part of wage and price dynamics. Two features are worth mentioning here. (For further details see the author, 1978.)

First the serial correlation structure for the output gap (or unemployment) in this model is hump shaped: the impact of shocks on output rises before diminishing toward zero. This hump shaped property is also present in the actual process for output or unemployment in the United States and a number of other countries. Hence, the model is capable of explaining not only the serial persistence of unemployment but also the shape of the persistence.

Second, a striking aspect of the *U.S.* quarterly data is that the humped shape reaches a peak at about one year; this corresponds to contract lengths in the model of about the same length. Hence, relatively short contracts (much shorter than the frequently cited three-year union contracts) are capable of displaying empirically observed serial persistence. Although other models might explain these correlations just as well, this type of model with relatively short contracts appears to be consistent with the data.

Whether the model is consistent with a rigorous micro theory is more difficult to determine. Unfortunately, the assumed contract formation behavior is not explicitly derived from a utility maximization model (see Robert Barro). While significant gains have been made in our understanding of contracts through the work by Costas Azariadis, Martin Baily (1974), and D. F. Gordon, the micro foundations of the staggered contract model presented here are far from complete. I think there are important information reasons for contracts to be staggered (without an auctioneer some staggering is necessary for firms to obtain information about the relative wage structure), but these are yet to be laid out rigorously. For this reason such models should be used cautiously since contract length is not a datum, and the optimal length of contracts may change with changes of policy. By way of comparison the information-based theories of aggregate dynamics which have been developed by Lucas also have problems with micro foundations. For example, disparate information is imposed on such models with little discussion of how stabilization policy might affect mobility or communication between markets which would alter this information structure.

The theoretical approach to micro foundations has proved difficult and is likely to remain so. But there is also an empirical approach which has received little attention. An early example of this approach is the

study of compensation policy made by Richard Lester in the late 1940's. His aim was to investigate (through a survey of firms) a number of alternative wage setting procedures: whether firms use wage surveys in determining their wage scales, whether firms try to anticipate future wage developments, and whether tight labor markets influence wage policy. While far from conclusive the study is suggestive of what might be done using modern techniques. For example, although wage surveys are now used almost universally by firms in determining wage scales, there is very little information available concerning how firms use these surveys. Such information would appear to be invaluable in modelling the macroeconomics of wage behavior.

IV. Concluding Remarks

The theme of this paper has been that the inflation dynamics typically associated with the "expectations-augmented" Phillips curve are significantly influenced by the interaction of staggered contracts as well as by expectations effects. While these ideas are implicit in much accelerationist research, the aim here has been to make them explicit in order that alternative hypotheses concerning the inflation process can be stated more clearly. The overlapping contract model described in the paper is closely related to a number of other models. (See George Akerlof, Baily, 1976, Fischer, Phelps, 1978, forthcoming, Stephen Ross and Wachter, and J. C. R. Rowley and D. A. Wilton, for example.) While the micro foundations of such models need to be developed more rigorously, they seem capable of improving our understanding of the dynamics of the inflationary process within a reasonable well-specified rational setting.

REFERENCES

G. Akerlof, "Relative Wages and the Rate of Inflation," *Quart. J. Econ.*, Aug. 1969, *83*, 353–74.

C. Azariadis, "Implicit Contracts and Unemployment Equilibria," *J. Polit. Econ.*, Dec. 1975, *83*, 1183–202.

M. N. Baily, "Contract Theory and the Moderation of Inflation by Recession and Controls," *Brookings Papers*, Washington 1976, *3*, 585–622.

———, "Wages and Employment under Uncertain Demand," *Rev. Econ. Statist.*, Jan. 1974, *41*, 37–50.

R. J. Barro, "Long-Term Contracting, Sticky Prices, and Monetary Policy," *J. Monet. Econ.*, July 1977, *3*, 305–16.

Philip Cagan, *The Hydra-Headed Monster. the Problem of Inflation in the United States*, Washington, 1974.

Rudiger Dornbusch and Stanley Fisher, *Macroeconomics*, New York 1978.

William J. Fellner, "The Core of the Controversy about Reducing Inflation: An Introductory Analysis," in his *Contemporary Economic Problems*, Washington 1978

———, *Towards a Reconstruction of Macroeconomics: Problems of Theory and Policy*, Washington 1976.

S. Fischer, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Feb. 1977, *85*, 191–205.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

D. F. Gordon, "A Neo-Classical Theory of Keynesian Unemployment," in Karl Brunner, and Allan H. Meltzer, eds., *The Phillips Curves and Labor Markets*, Amsterdam 1976, 65–97.

Robert J. Gordon, *Macroeconomics*, Boston 1978.

Richard A. Lester, *Company Wage Policies*, Princeton 1948.

R. E. Lucas, "Some International Evidence on Output Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

A. M. Okun, "Inflation: Its Mechanism and Welfare Costs," *Brookings Papers*, Washington 1975, *2*, 351–401.

G. L. Perry, "Slowing the Wage-Price Spiral: The Macroeconomic View," *Brookings Papers*, Washington 1978, *2*, 259–91.

E. S. Phelps, "Disinflation without Recession: Adaptive Guideposts and Monetary Policy," *Weltwirtsch. Arch.*, forthcoming.

———, "Introduction: Developments in Non-Walrasian Theory," in his *Studies in*

Macroeconomic Theory: Employment and Inflation, New York 1978.

———, "Money Wage Dynamics and Labor Market Equilibrium," in Edmund S. Phelps, et al., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

———, "Phillips Curves, Expectations of Inflation, and Optimal Employment over Time," *Economica*, Aug. 1967, *34*, 254–81.

S. A. Ross and M. L. Wachter, "Wage Determination, Inflation, and the Industrial Structure," *Amer. Econ. Rev.*, Sept. 1973, *63*, 675–94.

J. C. Rowley and D. A. Wilton, "Quarterly Models of Wage Determination: Some New Efficient Estimates," *Amer. Econ. Rev.*, June 1973, *63*, 380–89.

T. J. Sargent and N. Wallace, "'Rational' Expectations, the Optimal Monetary Instrument and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

J. B. Taylor, "Estimation and Control of a Macroeconomic Model with Rational Expectations," *Econometrica*, 1979.

———, "Aggregate Dynamics and Staggered Contracts," unpublished paper, Columbia Univ. 1978.

M. L. Wachter, "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers*, Washington 1976, *1*, 115–59.

Backward and Forward Solutions for Economies with Rational Expectations

By OLIVIER J. BLANCHARD*

In models where anticipations of future endogenous variables influence current behavior, there exists an infinity of solutions under the assumption of rational expectations. This problem has been dealt with, in the study of macro-economic models, by the implicit or explicit use of one of three additional requirements: optimality; consistency with alleged economic behavior; or conformity of the endogenous variables to an imposed stationarity condition. These requirements have coincided in existing models, leading to the choice of a unique solution, a "forward" solution. The purpose of this paper is to review the problem, characterize the solutions, and examine whether these requirements are acceptable. Section I presents a simple model and derives the set of solutions; the model makes no claim to generality, but has the major advantage that the issues are easily understood in this simple case. Section II discusses the requirement of consistency with economic behavior. Section III discusses the requirement of stationarity, and Section IV provides some conclusions.

I. The Model

Paul Samuelson's overlapping generation model with money is used. Agents live for two periods, receive one unit of perishable output in the first and can save only in the form of money. The government buys output from (sells output to) the young in exchange for money. Equilibrium is characterized by

$$\text{(1a)} \qquad (m_t - p_t)^d = -\alpha({}_tp_{t+1} - p_t)$$

$$\text{(1b)} \qquad (m_t - p_t) = (m_t - p_t)^d$$

$$\text{(1c)} \qquad {}_tp_{t+1} = E(p_{t+1} \mid I_t)$$

where m_t and p_t are the logarithms of the nominal money stock and of the price level, respectively; I_t is the information available to agents at time t, including current and past values of m_t and p_t; the notation ${}_{t-j}x_{t+i}$ denotes the agents' expectation of a variable x held at time t $-$ j for period t + i; and $E(.)$ denotes a mathematical expectation.

Equation (1a) states that the demand for money is a function of the expected rate of inflation. The sign of α is ambiguous: it depends on whether the elasticity of substitution between consumption in the two periods is less or greater than unity. If $\alpha > 0$, the equilibrium is formally equivalent to the equilibrium in the model of Philip Cagan. Equation (1b) characterizes market clearing: the demand for money is equal to the money stock inelastically supplied by the old and the government. Equation (1c) states that agents have rational expectations.

Combining these equations gives

$$\text{(2)} \quad p_t = \frac{1}{1+\alpha} m_t + \frac{\alpha}{1+\alpha} E(p_{t+1} \mid I_t)$$

A solution for p_t is simply a function:

$$\text{(3)} \quad p_t = \sum_{i=1}^{\infty} a_i m_{t-i} + bm_t + \sum_{i=1}^{\infty} c_i E(m_{t+i} \mid I_t)$$

with coefficients $(a_i)_i$, b, $(c_i)_i$ such that it satisfies (2). There are four remarks to be made:

The possibility that the price depends on other variables than money is excluded a priori. (This possibility has been examined by John Taylor and Robert Shiller.)

Variables such as past expectations of future money, formally $E(m_{t-j+i} \mid I_{t-j}), i, j > 0$ are omitted but not excluded a priori: if we included them, their coefficients would have to be equal to zero, for (3) to satisfy (2). This would not necessarily be the case in more general models.

Equation (3) allows past values of money

*Harvard University. I am indebted to Stanley Fischer, Benjamin Friedman, Edmund Phelps, Robert Solow, John Taylor, and Charles Wyplosz for useful comments and discussions.

to determine the current price level. Often the solution has been restricted by constraining $a = 0$ for all i, a priori.

The problem of convergence of the two infinite sums will be considered later. We may assume for the moment that there exists t_0 and t_1 such that $m_t = 0$ for $t < t_0$ and $E(m_{t-i}|I_t) = 0$ for $t + i > t_1$.

The coefficients are determined as follows: Leading equation (3) once and taking expectations on both sides, conditional on information available at time t:[1]

$$(4)\quad E(p_{t+1}|I_t) = \sum_{i=1}^{\infty} a_i m_{t-i+1} + bE(m_{t+1}|I_t) + \sum_{i=1}^{\infty} c_i E(m_{t+i+1}|I_t)$$

Replacing (4) in (2) and identifying term by term with (3), we can solve for the a_i and c_i as functions of α and b:

$$a_1 = \frac{b(1+\alpha) - 1}{\alpha};\ a_{i+1} = \left(\frac{1+\alpha}{\alpha}\right) a_i \qquad i = 1, 2, \ldots, \infty$$

$$c_1 = \frac{\alpha}{\alpha + 1} b;\quad c_{i+1} = \left(\frac{\alpha}{1+\alpha}\right) c_i \qquad i = 1, 2, \ldots, \infty$$

Therefore, there exists an infinity of solutions, each of them parameterized by b. Each solution can be written as a weighted average of two special solutions, a ***backward solution*** in which the price level depends on past values of the money stock ($b = 0$; $c_i = 0\ \forall\ i$):

$$(4')\qquad p_t^{(B)} = -\frac{1}{\alpha}\sum_{i=0}^{\infty}\left(\frac{1+\alpha}{\alpha}\right)^i m_{t-i-1}$$

and a *forward solution*[2] in which the price depends only on current and future expected values of m ($b = 1/(1+\alpha)$, $a_i = 0\ \forall\ i$):

$$(4'')\quad p_t^{(F)} = \frac{1}{1+\alpha} m_t + \frac{1}{1+\alpha} \cdot \sum_{i=1}^{\infty}\left(\frac{\alpha}{\alpha+1}\right)^i E(m_{t+i}|I_t)$$

Any solution p_t can be written as

$$(4''')\quad p_t = \lambda p_t^{(B)} + (1-\lambda)p_t^{(F)};\qquad \lambda \equiv 1 - b(1+\alpha)$$

Both the backward and the forward solution have been used in the literature. The backward solution is the solution traditionally used in the study of the dynamics of growth models. In the continuous time perfect foresight version of these models, it is referred to as the "myopic perfect foresight" assumption and can be obtained as the limiting case of adaptive expectations. The forward solution has been used in recent macro-economic models.

The indeterminacy does not depend on the use of discrete vs. continuous time or on certainty vs. uncertainty. The origin of the indeterminacy comes from the presence of an expected future value in the equilibrium equation. In each period both the current price and the expected future price clear the market. Over any number of periods, there is one more price (or expected price) than markets to clear. The indeterminacy will therefore be a general feature of models in which current prices depend on expected future prices (or the expected rate of change of prices). It is not present in models such as the no-speculation model of John Muth or the macro-economic model of Robert Lucas for example, in which only expected current values enter.

The indeterminacy is behaviorally significant: an unexpected increase in the nominal money stock, known to be permanent, leaves the real money stock unchanged under the forward solution, the price level unchanged this period under the backward solution. The sequence of utilities of agents will not be the same under different solutions.

We now consider the possibility of choosing between the solutions by imposing additional requirements. The first is the requirement of *optimality:* as the sequence of utilities depends on the solution chosen, it is likely that

[1]Use is made of the "law of iterated expectations":

$$E[E(m_{t+i+1}|I_{t+1})|I_t] = E(m_{t+i+1}|I_t)$$

[2]Edwin Burmeister R. Flood, and Stephen Turnovsky have remarked that the meaning of "forward" and "backward" conflicts with the meaning of the same words in the mathematics literature. They suggest the use of "forward looking" and "backward looking" would be less confusing.

the use of a given optimality criterion will lead to the choice of a unique solution. If agents were infinitely long lived, such as in Miguel Sidrauski, they would indeed choose the solution which maximizes their utility. In the model considered here however, agents live only for two periods and it is not clear what mechanism will lead to the choice of an optimal solution, however defined. Thus, other types of requirements have to be considered.

II. The Requirement of Consistency with Economic Behavior

The first argument was presented by Thomas Sargent and Neil Wallace, using the Cagan model. In the backward solution (my terminology), the price does not move in response to current changes in money; thus "next instant's price is what adjusts to insure equality between the demand and supply of real balances at this instant ... [whereas in the forward solution it is] the price level at each moment which adjusts instantaneously in order to insure that the real balances people hold equal the amount they would like to hold" (pp. 1044–45). Thus, they argue, the forward solution is more satisfactory.

Except for the pure backward solution however, both today's and expected next period prices move in response to a change in money: pinpointing which of the two clears the market is at best a difficult task. Furthermore, the interpretation given by Sargent and Wallace may not be the interpretation that agents have of the economy.

Consider the case where agents assume—rationally—that

$$p_t = \lambda^* p_t^{(B)} + (1 - \lambda^*) p_t^{(F)}$$

with λ^* close or equal to unity. If agents assume p_t to follow this process, they will be rational in believing that an increase in money this period affects next period's price level implying, depending on the value of α, expected inflation or deflation. This will lead them to increase their demand for real money balances; consequently, given the increase in the nominal money stock, a small change (or no change if $\lambda^* = 1$) in today's price equilibrates supply and demand. There is nothing in this description inconsistent with the way we think markets operate.

The second argument in favor of choosing the forward solution seems similarly flawed. It runs as follows: equation (2) only includes p_t, m_t, and $E(p_{t+1}| I_t)$ but no past variables. It is hard to see why the past should affect the current price at all. Only in the forward solution does the past not enter and thus this solution should be chosen. Equation (2) is however an incomplete description of the economy without an explicit expectation mechanism. If agents assume that p_t depends on the past in the way indicated by (4‴), then they will be rational and p_t will indeed depend on the past.

III. The Requirement of Stationarity

Heuristically, requiring stationarity amounts to requiring that if nominal money does not "explode," then the price level should not explode. More formally, if the logarithm of nominal money, m_t, follows a stationary process (with mean zero for convenience), so that it has a finite variance, then the equilibrium price level must also have finite variance. There are two separate questions: Why should such a requirement be imposed? Does imposing it lead to the choice of a unique solution? I consider them in turn.

Stationarity may follow from the requirements of optimality but, as indicated above, there is no reason why in this model the solution must satisfy optimality. Karl Shell and Joseph Stiglitz, and Edmund Phelps and Taylor have shown, however, that in certain models nonstationarity may violate the assumption of market clearing and of rationality of expectations. This also applies to this model and may be described—not rigorously—as follows. The model imposes implicitly a bound (upper or lower, depending on the value of α) on the expected rate of inflation. This expected rate cannot be such that it implies a demand for real money balances, or equivalently a supply of output by the young, larger than their endowment.

If p_t is nonstationary, then such an expected rate of inflation may be reached

with some positive probability. In this case either the market will not clear or the price will follow another solution, making expectations irrational. This may be a good reason therefore to require stationarity, without reference to optimality. But will this requirement lead to the choice of a unique solution? The answer depends on whether the elasticity of the current price with respect to the price expected next period is less or greater than unity in absolute value:[3]

A. If this elasticity is less than unity, i.e., if $|\alpha/(\alpha + 1)| < 1$ (or equivalently $\alpha > -1/2$), then only the forward solution is stationary. This follows directly from inspection of equations (4′) and (4″).

B. If this elasticity is greater than unity, i.e., if $|\alpha/(\alpha + 1)| > 1$, then clearly the backward solution is stationary. The forward solution may however also be stationary. Consider the following example:

$$m_t = \rho m_{t-1} + \eta_t, |\rho| < 1, \quad \eta_t \ IID$$

so that

$$E(m_{t+i} | I_t) = \rho^i m_t$$

If ρ is "small enough," i.e., if $|\alpha\rho/(\alpha + 1)| < 1$, then $p_t^{(F)}$ is stationary:

$$p_t^{(F)} = \frac{1}{1+\alpha} \sum_{i=0}^{\infty} \left(\frac{\alpha\rho}{\alpha + 1}\right)^i m_t$$

$$= \frac{1}{1 + \alpha(1 - \rho)} m_t$$

Therefore, if m_t is expected to return to its mean "fast enough," then the forward solution may be stationary. In this case all solutions are stationary. Requiring stationarity does not yield a unique solution. There are two ways in which the requirement of stationarity may be strengthened so as to yield a unique solution.

The first one is suggested by Taylor: it is to require that the price level not only have finite variance but also minimum variance. This criterion indeed allows us to choose a unique solution; one of its characteristics is that the choice is not independent of the stationary process followed by m_t.

The second one is suggested by the above example. If $|\alpha/(\alpha + 1)| > 1$, then the forward solution cannot be stationary for all stationary processes followed by m_t. In the above example, ρ can always be chosen so that $|\alpha\rho/(\alpha + 1)| > 1$. Thus, requiring that p_t be stationary for *any* stationary process generating m_t leads in this case to the choice of the backward solution.

Both of these strengthened criteria thus allow us to choose a unique (but possibly different) solution. However, they both lack the justification of the original stationarity criterion and it is hard to see why this decentralized economy will be led to apply them.

IV. Conclusion

Section I has characterized solutions as linear combinations of a backward and a forward solution. Section II has shown the requirement of consistency with alleged economic behavior to be unacceptable. Section III has shown that the requirement of stationarity may be justified but may not ensure the choice of a unique solution. This suggests two directions of research.

In this model, if the elasticity of the current price with respect to next period's expected price is less than unity, imposing stationarity leads to the choice of a unique solution—the forward solution. This raises two questions: how does this condition translate in more general models? Is such a condition likely to hold? Both questions are addressed in another paper (see the author). The answer is that the generalized condition is indeed likely to hold.

In models in which optimality cannot be invoked, and where imposing stationarity is not justified or does not lead to the choice of a unique solution, a new criterion must be found. A possible direction of research is the study of how the economy converges to a rational expectation solution and how "his-

[3]In this model, whether a solution is stationary or not depends on the value of a utility parameter α. Another approach would have been possible, allowing m_t to follow a simple feedback rule on p_t would have made the stationarity of a solution depend on the rule. The simplest example is $m_t = \gamma p_t$. Which solution is stationary depends on the value of γ. This is the approach followed by Fisher Black.

tory" may determine the value of λ. The difficulty lies in defining plausible revision rules. Using the revision rules suggested by Stephen Decanio, preliminary results indicate that, in the space of solutions, only the forward solution may be stable. If this result is robust, it might be the strongest argument for the choice of the forward solution.

REFERENCES

F. Black, "Uniqueness of Price Level in Monetary Growth Models with Rational Expectations," *J. Econ. Theory*, Jan. 1974, *7*, 53–65.

O. Blanchard, "The Solution of Linear Difference Models Under Rational Expectations; Its Application to the Hahn Problem," mimeo., Harvard Univ., June 1978.

E. Burmeister, R. Flood, and S. Turnovsky, "Rational Expectations and Stability in a Stochastic Monetary Model of Inflation," mimeo., Univ. Virginia, May 1978.

P. Cagan, "The Monetary Dynamics of Hyperinflation," in Milton Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

S. Decanio, "Rational Expectations and Learning from Experience," *Quart. J. Econ.*, forthcoming.

R. E. Lucas, Jr., "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

J. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

E. Phelps and J. Taylor, "Stabilizing Powers of Monetary Policy Under Rational Expectations," *J. Polit. Econ.*, Feb. 1977, *85*, 163–90.

T. Sargent and N. Wallace, "The Stability of Models of Money and Growth with Perfect Foresight," *Econometrica*, Nov. 1973, *41*, 1043–48.

K. Shell and J. Stiglitz, "The Allocation of Investment in a Dynamic Economy," *Quart. J. Econ.*, Nov. 1967, *81*, 592–609.

R. Shiller, "Rational Expectations and the Dynamic Structure of Macroeconomic Models," *J. Monet. Econ.*, Jan. 1978, *4*, 1–44.

M. Sidrauski, "Rational Choice and Patterns of Growth in a Monetary Economy," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 534–44.

J. Taylor, "Conditions for Unique Solutions in Stochastic Macro-Economic Models with Rational Expectations," *Econometrica*, Sept. 1977, *45*, 1377–85.

NEW DIRECTIONS FOR EMPLOYMENT POLICY

The Potential Impact of Employment Policy on the Unemployment Rate Consistent with Nonaccelerating Inflation

By GEORGE E. JOHNSON AND ARTHUR BLAKEMORE*

During the past few years there has been considerable interest, among both economists and politicians, in the question of how to lower the unemployment rate consistent with nonaccelerating inflation by the use of structural policies. Although it is not possible to estimate the value of this rate (which will hereafter be referred to as U_*) with any precision, its consensus estimate in 1978 terms is about 6 percent, plus or minus one-half of 1 percent. While recognition of this constraint on aggregate macro-economic policy has been taking hold, Congress has been debating a bill to establish a 4 percent overall unemployment rate as the national goal for 1983. Implicit in this bill (the so-called Humphrey-Hawkins proposal) is the necessity of expanding labor market programs in order to achieve its principal objective The major component of such programs in the United States has been the direct provision of job opportunities to specific groups in the labor force.[1] These generally take the form of public service employment (*PSE*), but there has also been some use of employment tax credits in the private sector.

There are several possible objectives of employment policy, ranging from fiscal relief for local governments to a substitution of subsidized employment for welfare. The purpose of this paper, however, is to examine the potential of employment policy to lower U_*, i.e., to shift the long-run Phillips curve to the left.[2] Can the provision of a large number of *PSE* jobs or the extensive use of wage subsidies alter the relationship between labor market tightness and the overall unemployment rate such that a lower unemployment can be sustained by macro-economic policy without accelerating inflation?

The answer to this question depends on how individual labor markets work—in particular, on the nature of unemployment among the persons toward whom the unemployment programs are targeted. There is little consensus among labor economists about the functioning of these markets and, accordingly, the question of the probable impact of employment policy on U_* is subject to considerable controversy. It is therefore useful to frame the issue in a manner which permits empirical testing. Such an attempt, as well as the associated empirical results, is presented in Section I of the paper. Some qualifications to these results, empirical and conceptual, are discussed in Section II.

I. A Simple Model

Because the impact of employment policy on U_* is the subject of this study, situations of

*University of Michigan, and Council on Wage and Price Stability, respectively. An initial draft of this paper was written while we were on the staff of the Council of Economic Advisers, and subsequent work on it was supported by Department of Labor Contract No. J-9-M-6-0009. The paper, alas, does not necessarily represent the official positions of any of these agencies. A longer version is available on request.

[1]The other four components of labor market policy included skill training, job search assistance, strengthening of work incentives, and the elimination of restrictive practices. Expenditures on employment programs, however, have dwarfed those on the other four.

[2]This should not be confused with the use of direct employment policy for countercyclical purposes, i.e., offering *PSE* jobs and/or wage subsidies during a recession and removing them when the economy recovers. The questions concerning the feasibility and desirability of employment policy for this purpose are very different from those concerning its use to lower U_*.

cyclical unemployment are not analyzed. Rather a very simplified equilibrium model of the aggregate labor market is developed in which the overall unemployment rate equals U_*. Two types of labor are identified: A workers who are relatively skilled (experienced, primary) and B workers who are unskilled (less experienced, secondary). Postponing specifics about the identification of A and B workers and the reason for their different unemployment situation, we will assert that the unemployment rate of the B's (U_b) is normally much higher than that of the A's (U_a). The employment level of B workers (E_b) relative to that of the A's depends negatively on the wage of the B's relative to the wage of the A's ($R = W_b/W_a$), or

(1) $E_b/E_a = q(R), dlog\, q/dlog\, R = -\sigma$

where σ is the elasticity of substitution between the two types of labor. For purposes of exposition we will assume that potential labor forces of both kinds of labor (L_a and L_b) are fixed, independent of both wages and job opportunities.

Employment policy operates by shifting the relative demand function for B workers to the right.[1] This can be represented as a subsidy to employers of s percent of their B-worker wage bill. Thus the relative cost of B's becomes $R(1 - s)$ and one additional *PSE* job for a B person implies that $ds = 1/\sigma E_b$.

The impact of employment policy on the aggregate level of employment and its distribution depends on how relative wages respond to shifts in relative demand. There are several approaches that can be pursued in this regard, but in the context of the model presented here it is useful to distinguish three possible descriptions of wage behavior. The relevance of these possibilities will be empirically tested. First, in what is called the *competitive case*, relative wages adjust completely to market forces. If the labor market operates in this manner, then the equilibrium unemployment rates of both A and B workers differ only because of turnover factors. For example, teenagers (who are B's) move in and out of the labor force much more frequently than prime age males (who are, by and large, A's), and this results in the much higher unemployment rate of teenagers (see, for example, Steven Marston). An increase or decrease in the employment demand for one group will cause vacancies that lead in turn to relative wage adjustments until the values of U_a and U_b are back, cyclical factors aside, to their original values of U_{a*} and U_{b*}. By this view of the labor market, employment policy can obviously have no influence on U_*.

The *minimum wage case* is the polar opposite of the competitive case. This time it is assumed that the wage of B workers is an institutionally fixed (not necessarily entirely by the government) fraction R_* of the wage of A workers. The labor market for A's, however, is the same as in the competitive case Thus, the employment of B's is given by $E_b = q(R_*(1 - s))(1 - U_a)L_a$, and the imposition of a subsidy on B labor will increase total employment and decrease U_*. In fact, one additional *PSE* job for a B worker will lower the level of unemployment by one.

There are several ways to justify wage behavior that yield results between the competitive and minimum wage cases. One way to do so, in the spirit of Martin Baily and James Tobin, is to assume that the equilibrium unemployment of each group depends negatively on its wage relative to the average wage in the economy. For this *relative unemployment case*, the unemployment rates are given by

$$(2) \quad U_i = \phi_i(W_i/W^*), \quad \frac{\phi_i' W_i/W^*}{1 - \phi_i} = -\epsilon_i$$

where W^* is the (weighted) average wage and ϵ_i is the elasticity of the ith group's employment rate with respect to W_i/W^*. This may be justified on the grounds that the rate of labor turnover of workers depends on their rate of compensation relative to the average. Note in the competitive case ϵ_a, ϵ_b equal zero, and in the minimum wage case ϵ_a equals zero

[1]This is unfortunately a theoretical assumption rather than a description of the actual administration of the employment programs of the past decade These have tended to benefit the middle of the skill distribution (see Johnson and James Tomola), which is equivalent to no shift in (1) at all. Recent policy changes have attempted to shift the emphasis of the programs toward the lower end of the skill distribution

and ϵ_b equals infinity.

Now, in the context of any of the above cases, the equilibrium condition in the labor market is

$$\frac{1 - U_b L_b}{1 - U_a L_a} = q(R(1 - s)) \tag{3}$$

Differentiating (3) totally with respect to R, L_b/L_a, and s and then plugging the result into the relative employment demand equation, we see that

$$(4)\quad d\log\left(\frac{E_b}{E_a}\right) = \frac{\sigma}{\sigma + \epsilon'}\, d\log\left(\frac{L_b}{L_a}\right) + \frac{\sigma\epsilon'}{\sigma + \epsilon'}\, ds$$

where $\epsilon' = \delta\epsilon_a + (1 - \delta)\epsilon_b$, δ being the weight of W_b in W^*.

There are two facts about (4) which are of importance. First, the equation theoretically reduces to any of the three wage descriptions and hence gives the respective employment policy impact. For a value of $\epsilon' = 0$ (the competitive case), employment policy has no influence on employment and supply creates its own demand. On the other hand, if the minimum wage case is correct ($\epsilon_a = 0$ and $\epsilon_b = \infty$), an increase in *PSE* jobs for B workers equal to x percent of E_b will increase the relative employment of B workers by x percent, but changes in the relative supply of B workers will have no influence on their relative employment. When ϵ' lies between zero and infinity, (the relative unemployment case) employment policy will have results intermediate between the polar cases.

The second fact is that equation (4) has very clear testable implications. The likely impact of future employment policy can be discerned by examining what happened to relative employment in response to past demographic shifts. The effect on the relative employment of B's due to a *PSE* program equal to x percent of their initial employment is $\epsilon'/(\sigma + \epsilon')$ times x, but $\epsilon'/(\sigma + \epsilon')$ is one minus the coefficient on $dlog(L_b/L_a)$ in (4). In order to infer the impact of ds on total rather than relative employment we require the values of both ϵ_a and ϵ_b, for $d\log E_a = -(\sigma\delta\epsilon_b/(\epsilon' + \sigma))ds$ and $d\log E_b = (\sigma(1 - \delta)\epsilon_b/(\epsilon' + \sigma))ds$. Thus, the change in aggregate employment in response to one additional *PSE* job for a B person is

$$(5)\qquad \Delta E = \frac{1 - \delta}{\epsilon' + \sigma}(\epsilon_b - \epsilon_a R)$$

This takes the value of zero in the competitive case, one in the minimum wage case, and between zero (assuming $\epsilon_b > \epsilon_a R$) and one in the relative unemployment case.

Empirical estimation of the parameters in (5) requires the use of a more realistic delineation of labor market types than that between our hypothetical A and B workers. Ideally, one would employ an extremely detailed breakdown by demographic and skill levels, but the appropriate data—as well as an adequate conceptual basis for making decisions on the appropriate classification—are not readily available. Therefore, as an *illustration* of the estimation of the relevant parameters we will use the working assumption that the labor market is segmented by demographic groups.[4] If we take two periods in which the overall unemployment rate was close to U_*, a cross-section regression of the between-period change in the logarithm of the *employment rate* of the ith demographic group on the between-period change in the logarithm of the average wage of that group will yield an estimate of the value of ϵ on the assumption that it is the same for all groups (in terms of the two-group model examined above, that $\epsilon_a = \epsilon_b$). A hypothesis that some set of demographic groups have a different ϵ_i than the rest of the labor force can be tested by adding an interaction term between the wage and a dummy variable for that set of groups.[5]

The results of regressions in which prime age (25 through 54 years of age) males are designated as the interaction term yield low

[4]A similar assumption is used in recent studies of the determination of the equilibrium unemployment rate by, for example, Michael Wachter and Robert Gordon.

[5]The observations are fourteen age/sex groups (16–19, 20–24, 25–34, 35–44, 45–54, 55–64, and 65+ for each of the sexes). The labor force data are from the Department of Labor. The wage data are constructed from full-time, year-around income from the Bureau of the Census. The latest available income data are 1976 figures, so this was extrapolated out one year for the 1977 observations.

estimates of ϵ. In other words there is evidence of a weak version of the relative unemployment model. When the between-period change in employment and wages is taken as the high employment years of 1970 and 1977, ϵ is .125 (.076 standard error) for prime age males, and .086 (.020) for the other groups in the labor force (males under 25 and over 55 and all women).[6] Thus, using these highly aggregated data there does appear to be a small statistically significant effect of the relative wage of "secondary" workers on the equilibrium unemployment rate of secondary workers. But for "primary" workers the estimated value of ϵ is neither statistically different from zero or the coefficient for secondary workers. For the period between 1956 and 1965, the estimated value of ϵ is $-.083$ (.147) for prime age males and .101 (.043) for the composite of the other groups. For the period between 1965 and 1970 the estimated coefficients are both small and insignificant, possibly because of the substantial change in the definition of unemployment that occurred in 1967.

The consensus of estimates of the elasticity of substitution between different types of labor is that it is greater than one but finite (see Daniel Hamermesh and James Grant). For the 1970–77 data, a regression of the change in the *log* of *the level of employment* of the *i*th group on the change in the *log* of its average wage yields an estimated slope coefficient of -1.43 (.44), which is a direct estimate of σ on the assumption that the elasticity of substitution between the different demographic groups is constant.[7]

Now consider a program that offers *PSE* jobs to persons in the labor force other than prime age males. The impact of such a program on the *level* of unemployment consistent with nonaccelerating inflation can be approximated by substituting the estimated values of ϵ for prime age males and others into (5). The values of δ and R in this case are approximately .4 and .55. Using the point estimates for 1970–77 (ϵ_a = .125 and ϵ_b = .086), the increase in total employment per *PSE* job is .006; with the higher point estimate of ϵ_b for 1956–65 (.101) and letting ϵ_a = 0, the increase in total employment per *PSE* job is .041.

II. Qualifications and Further Considerations

These results do not provide a very optimistic forecast of the potential of employment policy to lower the natural unemployment rate. The estimates imply that an additional one million *PSE* jobs for other than prime age males (at a cost of about $10 billion) would reduce the number of unemployed by 6,000 to 41,000. Even a doubling of the larger of these figures would mean a reduction of the overall unemployment rate of less than one-tenth of 1 percent.

This conclusion, however, may be overly pessimistic. The empirical estimates of the ϵ's were made at an extremely high level of aggregation, and it is possible (indeed, quite likely) that there are subsets of persons within these broad demographic groups for whom relative wages are not as flexible as for the rest of the group. In this case it would be *possible* for employment policy to have a much larger impact on U_* than that implied by our gross estimates so long as the programs were appropriately targeted toward those for whom the labor market "does not work." The leading candidate for such a group is minority teenagers; and indeed, over half of the resources for youth employment programs are targeted to minority youth. In other empirical work using more disaggregated data, we found that job programs for minority youth living in central cities and rural areas could increase total employment by approximately .4 per *PSE* slot. These results are also subject to qualifications. But for employment policy to be effective in reducing U_*, these targets

[6]These results are in principle subject to simultaneous equations bias because of the fact that changes in the employment rate, employment level, labor force participation rate, and relative wage of each group are jointly determined. However, the two-stage least squares estimate of ϵ is not perceptibly different from the ordinary least squares estimate reported above.

[7]This assumption is admittedly questionable for it means, for example, that teenage boys are equally substitutable for 45–54-year-old men as for teenage girls. The available econometric evidence suggests that such assumptions are not correct (see Richard Freeman), but the more general estimates are necessarily at a much higher level of aggregation than that used in these estimates.

must be clearly identified. The point of this paper is that—again, solely in terms of the objective of lowering U_*—identification of a set of characteristics of individuals that imply high unemployment rates is not a sufficient basis on which to target employment programs. In addition, one has to show that the removal from the labor force of X persons with those characteristics would not reduce significantly the total employment of persons with those characteristics.

Finally, there are several alternative conceptual approaches to the question of the equilibrium of the labor market, and these may have different implications concerning the effectiveness of employment policy. One attractive set of possibilities focuses on the effect of wage variation on the unemployment rate of individual groups in the labor force. Assume that any individual in each group can take a job with a lower, competitively determined wage or queue for a job with a higher, institutionally determined (by minimum wage legislation, union policy, or whatever) wage. Then there will be a component of the equilibrium unemployment rate of the group that depends positively on the relative disparity between these wages, either because of search considerations (see Jacob Mincer and Robert Hall) or because of labor supply effects[8] (a variant of a model developed by Finis Welch). In both of these models the effect of an employment program that attempts to expand higher and lower wage jobs in proportion to their initial base has an impact on the total employment of the group that bears a relationship to the effect of a labor force increase on employment analogous to equation (4). However, in the model that focuses on search unemployment, the provision of lower wage jobs is much more effective in reducing unemployment than the provision of high wage jobs (in fact, the latter may exacerbate the problem). In the model that focuses on labor supply effects, on the other hand, it is likely that an increase in the number of higher wage jobs would be more effective in reducing unemployment than an increase in the number of lower wage jobs.

[8]Specifically, some persons in the group would choose to participate in the labor force at the higher but not the lower wage, but since the number of higher wage jobs is restricted, they will report that they are unemployed even if lower wage jobs are available.

In any of these approaches, the results are not encouraging if the estimated parameters in this study are approximately correct. More disaggragated data are necessary to permit more reliable testing of the relative parameters.

REFERENCES

M. N. Baily and J. Tobin, "The Inflation-Unemployment Consequences of Job Creation Policies," in John Palmer, ed., *Creating Jobs: Public Employment Programs and Wage Subsidies*, Washington 1978.

R. Freeman, "Changes in the Age-Earning Profile and Substitution by Age," mimeo., Harvard Univ., Nov. 1977.

R. J. Gordon, "Structural Unemployment and the Productivity of Women," *J. Monet. Econ.*, Feb. 1977, suppl., *5*, 181–229.

R. E. Hall, "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers*, Washington 1975, *2*, 301–49.

D. Hamermesh and J. Grant, "Econometric Studies of Labor-Labor Substitution and Their Implications for Policy," mimeo., Michigan State Univ., June 1978.

G. E. Johnson and J. D. Tomola, "The Fiscal Substitution of Alternative Approaches to Public Service Employment," *J. Hum. Resources*, Winter 1977, *12*, 3–26.

S. Marston, "Employment Instability and High Unemployment Rates," *Brookings Papers*, Washington 1976, *1*, 169–210.

J. Mincer, "Unemployment Effects of Minimum Wages," *J. Polit. Econ.*, Aug. 1976, Part II, *84*, 87–104.

M. Wachter, "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers*, Washington 1976, *1*, 115–67.

F. Welch, "Minimum Wage Legislation in the United States," in Orley Ashenfelter and James Blum, eds., *Evaluating the Labor-Market Effects of Social Programs*, Princeton 1976.

U.S. Bureau of the Census, *Current Population Reports*, Series P-60, 1971; 1977.

U.S. Department of Labor, *The Employment and Training Report of the President*, Washington 1978.

Selective Employment Subsidies: Can Okun's Law Be Repealed?

By JOHN BISHOP AND ROBERT HAVEMAN*

Concern that structural factors impede efficient labor market performance is evidenced in both statistical analyses of economic potential and policy proposals for selective employment subsidies. Estimates of the level and expected growth of full-employment *GNP* have recently been revised downward, as has the 3.2 unemployment multiplier implicit in Okun's Law (see U.S. Council of Economic Advisers and George Perry). These indications of structural changes in labor markets reinforce statistics showing excessively high unemployment rates for youths and blacks, and labor force participation rates that are increasing for women and decreasing for men.

The simultaneous concern with high inflation and high measured unemployment, in the context of major changes in labor force composition and increased variance in sectoral unemployment rates (see Perry), has brought forth numerous and sizable selective employment subsidy policies (*SESP*) in both the United States and Western Europe. The *SESP*, changes in potential *GNP*, and Okun's Law are not unrelated phenomena. This paper explores that relationship. Section I presents a brief taxonomy of the primary *SESP*s which are currently being discussed in Western industrialized countries. Section II provides the economic rationale underlying these measures. Section III explores the relationship of *SESP* to the prospective growth of aggregate output, in the context of Okun's Law. Evidence on the existence and magnitude of changes in employment decisions in response to the New Jobs Tax Credit (*NJTC*) is presented in Section IV.

*Project associate, Institute for Research on Poverty, and professor of economics and fellow, Institute for Research on Poverty, University of Wisconsin-Madison, respectively. The helpful comments of John Palmer are acknowledged.

I

Wage (or employment) subsidies have been the primary measure designed to target employment demands on those sectors with substantial excess supply. They have appeared in various guises. A *SESP* can be a function of recruitment (additional hires), the existing employment stock, or changes in the employment stock. Each of these subsidies can be targeted on particular types of labor (say, by age, sex, region, unemployment duration, or education), or they can be general in nature. Moreover, the subsidy can be a flat amount or it can vary with the level of earnings, the wage rate, or the duration of coverage. It can be paid to the employer or to the worker, either directly or via a tax credit.

Examples of several of these variants have been recently implemented (see Haveman and G. B. Christainsen). The United States' *NJTC* is a constrained marginal stock subsidy with no targeting. In calendar years 1977 and 1978 firms expanding employment above 102 percent of the previous year's employment level receive a tax credit equal to 50 percent of the first $4,200 of wages paid each additional employee up to a maximum of 47 employees or $100,000 of credit. On the other hand, the 1975 British Temporary Employment Subsidy is a reverse recruitment rather than a stock subsidy, and like the *NJTC* it is temporary and nontargeted. This program subsidizes about 30 percent of the wage costs for up to one year of workers who would otherwise be laid off. In 1974, the West German government introduced a temporary targeted recruitment subsidy with a marginal stock constraint. For six months, a wage subsidy of 60 percent was paid to firms in specified regions for employing registered unemployed workers, if the firm's employ-

ment increased from that of a stipulated date prior to passage of the act.

The Netherlands, France, and Sweden have also recently adopted targeted employment subsidies. The percentage of the labor force on which *SESP*-type subsidies are paid varies from about .3 percent of the labor force in West Germany to 3–4 percent in Sweden. In 1978, the *NJTC* will be paid on the employment of nearly 1 percent of the *U.S.* labor force at a total budget cost of at least $2 billion

While few reliable evaluations have been made of these *SESP*s, the numerous extensions of what were to be temporary programs suggest that they have not been viewed as failures in achieving the primary objective—employment increases—set for them. In the United States, the Carter Administration has proposed replacing the lapsing *NJTC* with a Targeted Employment Tax Credit that would subsidize firms for 33 percent of the first $6,000 of first-year wages paid to low-income workers who are either 18–24 years of age or handicapped, and for 25 percent in the second year As revised in Congress, the proposal will likely become a hiring subsidy rather than a stock subsidy, and the target group will be expanded to include welfare recipients.

II

The economic rationale for *SESP* is straightforward: By reducing the price of labor at the margin, employment will be encouraged, unemployment reduced, price pressure will be reduced in competitive markets through a reduction in the marginal cost function for incremental output and, in the case of marginal stock subsidies, entry will be encouraged. Further, for firms engaged in external trade, *SESP* operates as an export subsidy (see Layard and Nickell). Indeed, for a number of Western European nations, this characteristic is viewed as a primary rationale for *SESP*. A temporary *SESP* encourages firms to incur labor costs earlier than otherwise. As a result, inventory accumulation or accelerated maintenance and investment spending will tend to increase. Finally, *SESP* (particularly nontemporary programs) will tend to induce the substitution of targeted labor for nontargeted labor. For example, it may induce adding a second shift rather than increasing overtime (see Jonathan Kesselman, Samuel Williamson, and Ernst Berndt).

Inevitably, however, the net job creation impact of a *SESP*—defined as the employment level in the economy with the policy less that without it—will, because of financial, output, and labor market displacements, be smaller than the gross number of jobs subsidized. A fully specified general equilibrium model is necessary to accurately estimate the net effects of a *SESP*.

If *SESP* is targeted on a resource in excess supply or with a positive and nontrivial supply elasticity (such as handicapped workers, transfer program recipients, and low-income youth (see Stanley Masters and Irwin Garfinkel)), potential *GNP*—defined as the level of *GNP* when *NAIRU* (the rate of unemployment that does not accelerate inflation) is attained—will rise. Even if the labor markets for these workers were free from distortions associated with minimum wage and tax and transfer programs, a wage subsidy of their employment paid for by a tax on other workers would raise potential *GNP* (see Bishop, 1977). Indeed, *SESP* can also increase potential *GNP* even if the labor force participation rate of each demographic group is fixed by reducing a wage-weighted *NAIRU* through concentrating employment increases on sectors with elastic sectoral Phillips curves (see Martin Baily and James Tobin).

The benefits of expanding potential *GNP* in this manner are increased by the fact that the labor supply decisions of targeted groups are distorted by high employer- and employee-paid taxes, and even higher marginal reduction rates of transfer benefits. Consequently, any resulting increase in actual and potential *GNP* through such employment increases will be positively correlated with the change in economic welfare. Moreover, pecuniary externalities for taxpayers are created by the increase in tax revenues and the decrease in transfer costs associated with *SESP*, both of

which reduce the net budgetary cost of the program.

Further a subsidy of one of the major costs of doing business will exercise downward pressure on prices during the transition to a new price level. This temporary reduction in inflationary pressure may feed back into inflationary expectations, and have a longer-run impact on price inflation.

In addition to its effects on actual and potential *GNP* and prices, *SESP* will tend to shift the composition of employment and earnings toward low-skill, target-group workers. If less inequality in the distribution of the adverse effects of poor economic performance is desired, this is a major benefit of *SESP*. One consequence of this redistribution is that, even with a constant *GNP*, the number of employed persons will increase as low-productivity workers are substituted for those with higher skills.

III

Because of these likely effects of *SESP*, the macro-economic relationships between changes in *GNP*, the *GNP* gap, and the unemployment rate will be altered. In standard policy models, increases in aggregate demand are viewed as closing the gap by increasing actual *GNP* toward some exogenously determined full-employment *GNP*. However, as indicated above, *SESP* is likely to increase simultaneously both the actual and a *NAIRU*-based potential *GNP*. Hence, a *SESP*-induced increase in *GNP* will reduce the conventionally measured *GNP* gap by more than it reduces the true gap. The *SESP* will also alter the relationship between the measured *GNP* gap and the unemployment rate. A *SESP*-induced increase in *GNP* will tend to be associated with a larger increase (decrease) in employment (unemployment) than is typically associated with general aggregate demand-induced changes in *GNP*.

Consider the following accounting relationship between *GNP*, productivity (A), employed capital (K), hours worked per week (H), and labor force participation rate (L):

$$(1)\quad dGNP = dA + (1 - B_L)\,dK + B_L(dH + S_n dL - S_n dU)$$

where dx indicates the percentage rate of change of x, $U = -100 \cdot log$(employment/labor force) $\approx$ the unemployment rate, B_L is the share of labor, and S_n is the ratio of the skill level of newly employed workers to the economy-wide average. Okun's Law is a reduced form of (1) which states that a 1 percentage point cyclical change in U is associated with a 3.2 percent change in *GNP*. While a percentage point decrease in U is directly associated in (1) with an increase in *GNP* equal to $B_L S_n$ (approximately .7 of a percentage point), cyclical changes in other determinants of *GNP*—namely, L, H, K, and A—are negatively associated with U. It is the sum of these effects that makes up the difference between .7 and 3.2.

There are at least three reasons why a 1 percentage point change in U induced by a *SESP* is not likely to increase *GNP* by 3.2 percentage points. First, a *SESP*-induced reduction in U will shift the composition of employment toward workers with $S_n < 1$ and increase the training costs of the firm. The inevitable result of such substitution is to reduce measured productivity, at least in the short run, and S_n and $|dA/dU|$ will fall as these costs are recorded in firm accounts.

Second, *SESP* encourages the hiring of part-time workers (especially if the per worker subsidized earnings level is capped) or the substitution of additional workers for increased overtime of existing workers. As a result, the response of H to changes in U will be smaller than otherwise—$|dH/dU|$ will fall.

Third, to the extent that target group labor is not complementary with capital services, as is likely, the utilization of capital will not increase as much as in the case of an equivalent general demand stimulus—$|dK/dU|$ will fall.

Finally, because of the limited knowledge on behavioral responses, the effect of *SESP* on $|dL/dU|$ is unknown. On the one hand, target group workers form a high proportion of the discouraged worker, nonlabor force

participant category. On the other hand, SESP may not generate as large an increase in labor force participation as an equivalent reduction in U stimulated by a general expansion in demand.

Thus, at least during the period of adjustment following enactment of a nontrivial SESP, Okun's Law is likely to be repealed. The reduction in the Okun unemployment multiplier associated with *SESP* is evidence that the policy is having the effects for which it is designed—increasing potential *GNP* and redistributing the costs of unemployment.

However, these effects do not come at zero cost. The *SESP* is not easy to administer. A marginal stock variety of *SESP* tends to favor new and fast-growing firms and regions. In addition, *SESP* may increase labor turnover, especially if it is temporary or of the recruitment variety. Finally, *SESP* with narrowly defined target groups (for example, low-income youth or welfare recipients) may result in the displacement of equally disadvantaged workers who may have more central positions in family units.

IV

None of these impacts of *SESP* will materialize if firms fail to change their behavior in response to the subsidy. In some past programs (for example, *WIN* and *JOBS*), that response was not substantial (see Daniel Hamermesh). Administrative costs or the low-productivity signaling effect of the subsidy apparently weakened the employment incentive for which the programs were designed.

The *NJTC* has been in operation for more than a year. While a definitive assessment of its effect on employment and prices is not yet possible, a preliminary evaluation can be made. In theory, the *NJTC* should provide a major stimulus to employment, as firms which typically hire part-time or part-year workers will find that the labor costs of an expansion are cut nearly in half. The \$100,000 per firm limit on the subsidy suggests that small and medium sized firms will experience the largest employment incentive.

Nonseasonally adjusted monthly data on employment and man-hours in construction and retailing were regressed on seasonal dummies, trends on the seasonal dummies, and three-year distributed lags of input prices (gross employer wages, W; wholesale price of construction materials, M, or consumer finished goods, P; materials, services, and energy prices, Q; gasoline and electricity prices, G; and a rental price of capital, R). The lag structures were freely estimated, with each input price or price ratio being represented by its contemporaneous value, and that of each of the previous four quarters and four half-years. Exogeneity tests which entered future values of the wage rate into the equation tended to confirm the hypothesis that wage and man-hours are simultaneously determined. Consequently, all models were estimated using two-stage least squares.

The *NJTC* variable is an average over the past six months of the proportion of firms (weighted by employees) that had knowledge of the credit. It has a value of .057 in June 1977 and rises at an average rate of .0424 per month, reaching .343 in January 1978 and .572 in June 1978

All of the *NJTC* coefficients are positive and significant in Models I and II, where input prices enter as ratios (see Table 1). When input prices enter nominally (Models III and IV), the coefficients are smaller and insignificant. Across all of the regressions the (point) average *NJTC* employment stimulus over the mid-1977 to mid-1978 period ranges from 150,000–670,000. For these industries, actual total employment growth over the period was 1.3 million. The Model III and IV estimates attribute at least 20–30 percent of the observed employment increase in these industries to *NJTC*. These results are consistent with the observation that between 1977II and 1978II rates of employment growth have substantially exceeded rates of output growth in both construction and retailing. In construction, employment grew at an 8.2–9.9 percent rate—double the 4.5 percent growth rate of real construction output. Even in

TABLE 1—IMPACT OF THE *NJTC* ON EMPLOYMENT IN CONSTRUCTION AND DISTRIBUTION

Sample—1952–78:06	Coefficient on *NJTC* under Alternative Specifications[a] I	II	III	IV
Employment				
Wholesale and Retail	.076[b]	.101[c]	.064	.061
Household Data	(.048)	(050)	(.059)	(.044)
	0121	*.0119*	*.0121*	*0122*
Retail	.045[c]	.047[c]	.030	.031
Establishment Data	(.019)	(.020)	(.023)	(.020)
	0044	*.0045*	*.0044*	*.0050*
Construction	196[c]		.166	.052
Household Data	(079)		(108)	(102)
	0336		*0297*	*.0353*
Construction	180[c]		.020	.052
Establishment Data	(053)		(.108)	(.102)
	0175		*.0155*	*.0173*
Man-Hours				
Construction	110		.025	.007
Establishment Data	(.078)		(108)	(.091)
	0340		*0302*	*.032*
Average *NJTC*-Induced Employment Δ in 12-month period preceeding June 1978 (in thousands)				
Household Data	669	575	565	410
Establishment Data	412	203	154	255

Source: See Bishop (1978).

[a]Estimated with two-stage least squares The standard error of the coefficient is shown in parentheses, and the standard error of the estimate in italics beneath the coefficient. Model I $E = \beta_0 \cdot NJTC + \beta_1 X + \beta_2(W/P) + \beta_3(R/P) + \beta_4(Q/P)$ for retailing and $E = \beta_0 \cdot NJTC + \beta_1 X + \beta_2(W/M) + \beta_3(R/M)$ for construction where X is the vector of output lags, seasonal dummies and trends Model II Adds β_5 (G/P) to Model I Model III. Enters W, R, Q, M, and P nominally, rather than as ratios. Model IV: Same as III, but with distributed lags limited to 1.5 rather than three years

[b]Significant at 05 level on a one-tail test

[c]Significant at 025 level on a one-tail test.

retailing, where cyclical changes in employment are small, the 3.0 percent growth of real sales lagged behind the 3.4–4.0 percent growth of employment. The contrast between construction man-hours and employment regressions also suggests that the *NJTC* has, as predicted, caused a reduction in average hours per week (see Bishop, 1978).

To test the relationship between prices and subsidy-induced marginal cost reductions, the monthly change in retail price was regressed on current and lagged changes in a number of industry cost variables (wages, wholesale product prices, materials, services, and energy prices, a rental price of capital, and excise taxes), the unemployment rate, and the level and trends in seasonal dummies. For nonfood commodities and restaurant meals the retail trade margin is negatively and significantly related to the *NJTC* variable (see Table 2). Between May 1977 and June 1978, retail nonfood commodity prices rose 4.73 percent, while the counterpart wholesale prices rose 6.56 percent. This discrepancy of 1.83 percentage points approximates the preferred *NJTC* estimated effect of 2.2 percent (.038 · .572 · 100) (col. 1). The observed decline in the margin is particularly surprising given

TABLE 2—IMPACT OF THE *NJTC* ON THE MARGIN BETWEEN RETAIL AND WHOLESALE PRICES

	Coefficient on *NJTC* under Alternative Specifications[a]			
	One-Year Distributed Lag			Six-Month Lag
	Trends on Seasonals		No Trends	Trends
CPI Component	with Q	without Q	with Q	with Q
Food Away from Home	−.036[c]	−.037[c]	−.032[c]	−.033[c]
	(013)	(.012)	(.013)	(.013)
	0017	*.0017*	*.0017*	*.0018*
Nonfood Commodities	−.038[c]	−.038[c]	−.031[b]	−.038[c]
	(.015)	(.015)	(016)	(.015)
	.0020	*0021*	*.0022*	*.0020*
Food at Home	.051	041	051	.051
	(039)	(.038)	(.040)	(.038)
	0053	*.0053*	*.0052*	*.0052*
All Commodities	−.018	−.019	− 013	−.018
	(.016)	(.016)	(.017)	(016)
	0022	*0022*	*.0023*	*.0022*
Reduction in Consumer Costs between June 1977 and June 1978 (in billions)				
All Commodity Regression	3 4	3 6	2.4	3.4
Disaggregated Regressions	2.8	3 3	1.9	2.8

Source: See Bishop (1978).

[a]The standard error of the coefficient is shown in parentheses and the regression is shown in italics beneath the coefficient Models estimated on monthly data 1953:03 to 1978.06. Weights for Q are based on the 1967 input-output table, which includes gasoline, electricity, telephones, containers, cellophane packaging, supplies, insurance, auto repair, and legal fees.

[b]Significant at 05 level on a one-tail test

[c]Significant at 025 level on a one-tail test

recent increases in the relative price of imported consumer goods. (Imported products, it should be noted, are included in retail but not wholesale price indexes.)

Among the subsectors, the pattern of coefficients is consistent with expectations. The large negative coefficients for the low-skill-intensive restaurant industry suggest that the 8–12 percent induced reduction in marginal costs caused a 1.1 percent decline in output price. On the other hand, the small margin, nonwage-intensive retail food industry has a nonsignificant positive coefficient, reflecting the greater contribution of incremental employment in this sector to quality than to the volume of output.

The final rows of the table indicate that the reduction of consumer costs attributable to the *NJTC* ranges from $1.9–3.6 billion. By comparison, it is predicted that over its two-year life, *NJTC* credit claims will be $3–6 billion.

These estimates of the impact of *NJTC* are for those sectors with the largest expected response. While across-industry displacements might offset these impacts, there is no clear reason why this would occur. While limited awareness of *NJTC* may have reduced its measured effectiveness, its temporary character may have led to inventory

accumulation which a more permanent program would not induce.

No impact estimates based on only the first year of program experience can be conclusive. Perhaps the *NJTC* variable is capturing other exogenous forces inducing contemporaneous employment increases and price decreases, in which case improved specifications may reduce the estimated *NJTC* impacts. While a number of possibilities have been tried (for example, varying lags, trend shifts in 1974, and the addition of an energy price variable) without significant effects on the *NJTC* variable, other factors may be at work. Hence, the finding that *NJTC* has had sizable employment and price effects must remain tentative. It should be noted, however, that the procedure employed is more robust with respect to assumptions on the impact of taxation changes than those used to estimate the response of investment spending to taxation changes.

V

In sum, the case for *SESP* is a strong one. The level of employment is likely to be increased, its composition improved, *NAIRU* reduced, and the associated price increase lower with *SESP* than with an equivalent general stimulus to aggregate demand. And the (at least temporary) reduction in the Okun multiplier is evidence that *SESP* is inducing the behavior for which it was designed. The results from the *NJTC* regressions suggest that such employer hiring and price responses do occur. However, these responses are for a nontargeted program; extrapolation of them to a targeted *SESP* would be inconsistent with evaluations of previous such programs.

REFERENCES

M. Baily and J. Tobin, "Inflation-Unemployment Consequences of Job Creation Policies," in John Palmer, ed., *Creating Jobs. Public Employment Programs and Wage Subsidies*, Washington 1978.

J. Bishop, "The General Equilibrium Impact of Alternative Anti-Poverty Strategies: Income Maintenance, Training, and Job Creation," *Ind. Labor Relat. Rev.*, forthcoming.

———, "Pricing and Employment in Construction and Distribution: The Impacts of Tax Policy," paper presented at Universities-Nat. Bur. Econ. Res., Conference on Low Income Labor Markets, June 9–10, 1978.

D. Hamermesh, "Subsidies for Jobs in the Private Sector," in John Palmer, ed., *Creating Jobs: Public Employment Programs and Wage Subsidies*, Washington 1978.

R. H. Haveman and G. B. Christainsen, "Public Employment and Wage Subsidies in Western Europe and the U.S.: What We're Doing and What We Know," National Commission on Manpower Policy, forthcoming.

J. Kesselman, S. Williamson, and E. Berndt, "Tax Credits for Employment Rather Than Investment," *Amer. Econ. Rev.*, June 1977, *67*, 339–49.

Stanley Masters and Irwin Garfinkel, *Estimating the Labor Supply Effects of Income Maintenance Alternatives*, New York 1978.

G. L. Perry, "Potential Output and Productivity," *Brookings Papers*, Washington 1977, *1*, 11–47.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, Jan. 1977, 52–56.

Retirement Policies, Employment, and Unemployment

By RONALD G. EHRENBERG*

There is a growing consensus among economists that reliance on aggregate demand policies alone will not be sufficient to move the economy to full employment with a nonaccelerating inflation rate, and that policies which alter the structure of labor markets will be required. While obvious structural policies such as public sector employment programs and training programs are the focus of current debate, many other public policies affect labor markets in subtle ways which may well adversely affect the level and distribution of employment and unemployment. To help improve the inflation-unemployment tradeoff, policymakers should seek to marginally modify these policies, preserving their benefits while reducing their adverse labor market effects.

To illustrate these points, this paper discusses the influence of public and private retirement policies on the level and distribution of employment and unemployment. I focus on the Social Security system (*OASDHI*), the Employee Retirement Income Security Act (*ERISA*), the amendment to the Age Discrimination in Employment Act that raised the permissible mandatory retirement age to 70, the Supreme Court decision in the Manhart case prohibiting sex differentials in employee pension contributions, and early retirement provisions negotiated in private collective bargaining agreements. Certainly, it would be difficult to criticize the *intent* of these policies. However, each of the *public* policies adversely affects the level or distribution of employment and unemployment. I conclude by noting several reforms of the method of financing the Social Security system which would reduce the system's adverse labor market effects.

*Professor of economics and labor economics, Cornell University. Support for my research was provided by NSF Grant No. SOC 77-15800.

I. The Social Security System

The Social Security system influences labor markets in a variety of ways. First, the *retirement earnings test* for receipt of benefits and the 50 percent *marginal tax rate* on earnings above the $4,000 *earnings exemption* discourage labor force participation and employment of the aged (see Michael Boskin). These parameters, by reducing the net return to work effort after age 65, also induce a life cycle reallocation of work effort from the retirement years to earlier years (see James Smith). Empirical evidence suggests that the work week of prime age males may have *increased* by over two hours above the level it otherwise would have been because of this effect (see Richard Burkhauser and John Turner).

Second, if employers cannot shift 100 percent of the share of the payroll tax paid by them onto employees in the form of lower wages (or smaller wage increases), then firms' employment decisions will be affected. Although evidence on the extent of shifting is mixed, two recent studies concluded that less than 50 percent of employers' share of the tax is shifted onto labor (see Daniel Hamermesh, 1977a; the author, Robert Hutchens, and Robert Smith), which should induce employers to hire fewer employees than they would in the tax's absence.

Furthermore, the existence of a maximum taxable earnings level causes payroll tax rate increases to increase the cost of low-wage employees relative to the costs of high-wage employees. If relative wages do not fully adjust, increases in the tax rate should lead firms to substitute high-wage for low-wage workers. In contrast, increases in the taxable earnings level reduce the incentives for such substitution (see John Pencavel). Between 1960 and 1978, the *OASDHI* tax rate more than doubled, while the maximum taxable

earnings level rose from $4,800 to $17,700. The latter change has likely dominated the former, causing a reduction in employers' incentives to substitute high-wage for low-wage employees and a shift in the distribution of employment (unemployment) towards (away from) low-skilled individuals.

The share of the payroll tax either nominally paid by employees or implicitly paid by them in the form of lower wages has a differential impact on different classes of individuals. For individuals outside the labor force it has a pure substitution effect, discouraging labor force participation. For employed individuals earning more than the taxable earnings level, it has a pure income effect, stimulating increased work effort. For employed individuals earning less than the taxable wage base, both effects are present, and the net impact is ambiguous.

The large increases in both the tax rate and maximum taxable earnings levels during the past decade may have reduced the work effort of individuals who earned more than the maximum taxable level prior to the increase, but less after. Although the impact of these changes on the unemployment rate is ambiguous, their net effect was probably to reduce the growth rate of employment. This effect may have been partially offset by the accompanying liberalization of promised future benefits. Since eligibility for *OASI* benefits depends upon career work effort, promised higher future benefit levels may stimulate greater work effort on the part of nonaged workers. However, this *entitlement effect* is likely to be greatest for low-wage workers as the benefit-earnings ratio declines as earnings rise. Moreover, since married females have the option of receiving either their own benefits or 50 percent of their husband's benefits (100 percent after he dies), their lifetime work effort entitles them to only small net additional *OASI* benefits and the entitlement effect is likely to be unimportant for them.

Finally, recent evidence suggests that the Social Security system may substantially reduce private savings (see Martin Feldstein, 1974a,b 1976; Alicia Munnell, 1974; and for contrasting evidence, Robert Barro). This net reduction is *not* offset by an increase in publi savings because of the pay-as-you-go natur of the system. As a result, total savings an capital accumulation in the economy ar reduced, leading to reduced growth in produc tivity and output, and ultimately to reduce rates of growth of employment and/or rea wages. Recent increases in Social Securit taxes and promised future benefits levels hav likely exacerbated this effect.

In sum, the parameters of the system inter act to produce numerous effects on labo markets. The reduction in labor force partici pation and employment of the aged is planned effect and should not be judged negative feature. In contrast, the *OASDH* payroll tax on employers *and* employees an the unfunded nature of the retirement trus fund probably serve to reduce both the labo force participation rates and employmen levels of the nonaged. The parameters of th system also differentially influence the distri bution of employment and unemploymen across sex classes and earnings classes o employees. Recent changes in the system' parameters probably have marginally slowe the growth rate of employment and reduce employers' incentives to substitute high- fo low-skilled labor.

II. Employee Retirement Income Security Act of 1974 *(ERISA)*

The *ERISA* was designed to increase th probability that private sector employee receive promised retirement benefits. I includes provisions requiring liberalized vest ing rules, more stringent funding require ments, and increased fiduciary responsibilit These provisions increase employers' costs c providing pensions and should lead employer to shift at least part of the increased costs t employees in the form of lower wages, smalle wage increases, or pension plan termination Although it is too early to assess *ERISA*' impact on wages, recent studies show that tradeoff exists between wages and retiremen system characteristics in both the private an public sectors (see the author; Alan Gustma and Martin Segal; Randall Weiss and Brad ley Schiller). If employers cannot fully shi

ERISA's costs, unit labor costs will increase, resulting in a reduction in the level (or rate of growth) of private employment. This reduction would be concentrated in those firms with pension plans whose pre-*ERISA* provisions did not meet the *ERISA* standards. Adoption of *ERISA*-type controls over public employees' retirement systems would have a similar negative impact on employment in the public sector.

The *ERISA*-type controls also affect the level of pension plan funding and composition of pension funds' portfolios. By requiring pension plans to be fully funded, they increase the stock of curent pension fund assets which, if not offset by a decline in individuals' saving, will increase the level of capital accumulation and ultimately the level of employment. On the other hand, by restricting the type of investments which pension funds may make, the controls prevent pension fund assets from being invested in projects with the highest expected rate of return (but also highest risk) and hence reduce the rate of productivity growth. Without empirical evidence, one can not ascertain which of these effects dominates.

III. Mandatory Retirement

The amendment to the Age Discrimination in Employment Act passed by Congress earlier this year, subject to a few exceptions, raises from 65 to 70 the age at which employers may compel their employees to retire. This will influence the level and distribution of employment in a number of ways. Mandatory retirement provisions tend to be found in large establishments which are unionized and in which employees usually have long actual or expected job tenure (see Edward Lazear). The typical life cycle relationship between an individual's earnings and productivity in firms, where an implicit long-term contract exists between the firm and its employees, is one in which earnings first exceed productivity, then productivity exceeds earnings, and finally earnings again exceed productivity. These stages correspond to a period of formal or informal training, a period of peak productivity, and a final period in which productivity is declining, but informal rules or union contracts prevent wages from being cut. The age at which this latter period starts, if at all, varies widely across individuals and depends upon factors such as the employee's health and the demands of his or her specific job. The establishment of a mandatory retirement age at an age such that, *on average*, the present value of employees' earnings just equals the present value of their productivity allows a firm to maximize its expected presented value of its profits. Such rules also allow increases in the present value of employees' earnings over their life cycles (see Lazear).

If the legislation induces some individuals to postpone their retirement, then on average the present value of wages will exceed the present value of marginal productivities over employees' careers. Employers may respond by negotiating flatter or everywhere lower real wage profiles. The overall *level* of employment would be unchanged, however new hires would be reduced, because the average employee would have a longer work-life. Hence, some jobs would be redistributed from new hires, primarily youths, to the aged. Further if employers face any difficulty in making wage adjustments, then they will tend to reduce their *stock* of employees, causing still *larger* reductions in new hires.

The change may also discourage employers from hiring middle-aged employees. Prior to the legislation, a firm would be willing to hire a middle-aged worker provided that his expected present value of marginal productivities less wages was nonnegative. If expected wages at the old retirement age exceeded expected marginal productivities, the legislative-induced increase in the expected retirement age reduces the firm's incentive to hire middle-aged workers and the maximum age at which it will hire new employees. Indeed, this provides employers an added incentive to prefer young rather than middle-aged new hires, and partially offsets the legislation's negative impact on youth employment (see Barry Chiswick and Carmel Chiswick). However, to the extent that the legislation reduces the number of retirees per year, employers may be forced to increase layoffs to

achieve desired lower employment levels in periods of declining aggregate demand, causing a further redistribution of employment away from those with the least seniority and increasing the measured unemployment rate.

The magnitudes of all of these effects depend upon the number of retirements postponed in response to the legislation; one recent study concluded some 200,000 aged employees would be added to the work force in the first year (see U.S. Department of Labor). However, growth in real incomes, private pensions, and Social Security benefits have reduced males' average age at retirement and as long as the Social Security retirement earnings test rules are maintained, workers aged 65 face a substantial incentive to retire. Thus, although the legislated change may marginally alter the distribution of employment and unemployment across age groups, its overall effect on the level of employment is likely to be small. It may, however, also substantially slow the progress of nonwhites into professional positions (see George Johnson and Juli Malveaux).

IV. The Manhart Case

On April 25, 1978, the U.S. Supreme Court declared that employers who require females to contribute a greater proportion of their salaries than males to contributory pension plans are committing illegal sex discrimination (see City of Los Angeles vs. Manhart). However, because female life expectancies are longer than males', to maintain the actuarial soundness of a defined benefit pension plan females must either (a) receive lower annual retirement benefits than otherwise identical males, or (b) receive equal annual retirement benefits, with larger annual contributions being made for females. The Supreme Court decision prohibits (b) unless the larger contribution is nominally paid by employers. This increases the relative costs of female employees, providing employers with an incentive to substitute males for females.

One proposal to eliminate this incentive is to use a "unisex" mortality table, calculated by weighting the relevant male and female mortality tables by the proportion of employees of each sex employed by a firm. Equal net contribution rates for *all* employees of a given age necessary to fully fund equal retirement benefits per year for retirees of each sex could then be determined. However, employers should realize that by reducing the proportion of females in their work force, they would reduce their required average net contributions (see Burt Barnow and the author). The likely magnitude of this substitution depends upon the true pension cost differential between males and female employees, and the extent to which males and females are substitutes in production. The former is likely to be quite small in plans which provide survivors benefits for spouses of beneficiaries, while precise estimates of the latter have yet to be obtained (see Hamermesh and James Grant).

V. Early Retirement

Early retirement provisions contained in many privately negotiated contracts typically allow early retirement at reduced benefit levels. While early retirement provisions are of value to employees, they also have the effect of redistributing employment losses across age groups of employees during periods of low or declining demand which may well *reduce* employers' costs.

Union contracts typically require that layoffs be inversely related to seniority; however utilizing such a policy to reduce employment may not be optimal from an employer's prospective. Due to the experience-rated nature of the unemployment insurance (*UI*) payroll tax, after some point layoffs raise the employer's payroll tax. Moreover, if the firm's most senior workers are in the stage of their life cycles in which wages exceed marginal productivities, the firm would best be served by reducing their employment rather than younger workers. Furthermore, if these senior workers voluntarily leave their jobs, an employer's *UI* tax rate would not increase as voluntary separations are not eligible for *UI* benefits in most states. Early retirement provisions thus allow employers to

redistribute employment losses in periods of low or declining demand from younger to older workers and to reduce their *UI* payroll tax contributions (see James Medoff). Since retirees tend to be out of the labor force, these policies probably do reduce the measured unemployment rate.

VI. Conclusion

All of the retirement policies discussed in this paper, except for privately negotiated early retirement provisions, were shown to have adverse effects on the level and distribution of employment and unemployment. These examples support the contention that more explicit attention should be given to the employment effects of social programs prior to their adoption *and* that consideration should be given to restructuring existing programs to reduce their adverse labor market effects. While each of the effects is probably quite small, their sum may be sizable.

Three examples of possible changes in the financing of the Social Security system illustrate the types of restructuring one might consider. First, the use of general revenue financing from personal and corporate income tax revenues, for all or some fraction of future system revenue needs, would reduce employer's incentives to substitute capital for labor. Second, increasing system revenues by more than is necessary to fund benefits in the short-run to build up a larger Social Security trust fund, and using this fund to buy outstanding government debt, would increase the social rate of savings and capital accumulation which ultimately would result in increased rates of growth of employment (see Feldstein, 1977). Third, raising the maximum taxable earnings level, rather than the payroll tax rate, to meet future system revenue needs would reduce employers' incentives to substitute high-wage for low-wage workers. To the extent that the overall rate of wage inflation is influenced more by the level of excess demand for labor in high-wage labor markets than that in low-wage labor markets, this change will also reduce the unemployment rate associated with each level of inflation (see Martin Baily and James Tobin; George Johnson and Arthur Blakemore).

REFERENCES

M. N. Baily and J. Tobin, "Macroeconomic Effects of Selective Public Employment and Wage Subsidies," *Brookings Papers*, Washington 1977, *2*, 511–44.

B. S. Barnow and R. G. Ehrenberg, "The Costs of Defined Benefit Pension Plans and Firm Adjustments," *Quart. J. Econ.*, forthcoming.

R. Barro, "Social Security and Private Saving: Evidence from the U.S. Time-Series," mimeo, Univ. Rochester 1977.

M. Boskin, "Social Security and Retirement Decision," *Econ. Inquiry*, Jan. 1977, *15*, 1–25.

R. V. Burkhauser, and J. A. Turner, "A Time-Series Analysis on Social Security and its Effect on the Market Work of Men at Younger Ages," *J. Polit. Econ.*, Aug. 1978, *86*, 701–15.

B. R. Chiswick and C. U. Chiswick, "On Benefits of Mandatory Retirement," *New York Times*, Nov. 12, 1977.

R. G. Ehrenberg, "Retirement System Characteristics and Compensating Differentials in the Public Sector," paper presented to the Econometric Society Meetings, Chicago 1978.

———, **R. Hutchens, and R. S. Smith,** "The Distribution of Unemployment Insurance Benefits and Costs," U.S. Department of Labor, Final Report, Contract J-9-M-6-0098, Mar. 1978.

M. Feldstein, (1974a) "Social Security, Induced Retirement and Aggregate Capital Accumulation," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 905–26.

———, (1974b) "Social Security and Private Savings: International Evidence in an Extended Life Model," in his *The Economics of Public Services*, New York 1974.

———, "Toward A Reform of Social Security," *Publ. Interest*, Summer 1975, *40*, 75–95.

———, "Social Security and Saving: The Extended Life Cycle Theory," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 77–86.

———, "Facing the Social Security Crisis," *Publ. Interest*, Spring 1977, *47*, 88–100.

A. Gustman and M. Segal, "Interstate Variations in Teachers' Pensions," *Ind. Relat.*, Oct. 1977, *16*, 335–44.

D. Hamermesh, (1977a) "New Estimates of the Incidence of the Payroll Tax," mimeo., Michigan State Univ. Aug. 1977.

———, "Effect of the UI System on Labor Force Behavior," tech. anal. paper no. 54, U.S. Department of Labor, Sept. 1977.

——— **and J. Grant,** "Econometric Studies of Labor—Labor Substitution and Their Implications for Policy," paper presented at the Allied Social Sciences Meetings, Chicago, Aug. 1978.

G. Johnson and A. Blakemore, "Estimating the Potential for Reducing the Unemployment Rate Consistent with Non-Accelerating Inflation: Methodological Issues," mimeo., U.S. Council Econ. Advisors, Mar. 1978.

——— **and J. Malveaux,** "Mandatory Retirement and Affirmative Action," mimeo., U.S. Council Econ. Advisors, Aug. 1977.

E. Lazear, "Why is There Mandatory Retirement?," mimeo., Univ. Chicago, Nov. 1977.

J. Medoff, "Layoffs and Alternatives under Trade Unions in U.S. Manufacturing," *Amer. Econ. Rev.*, forthcoming.

Alicia Munnell, *The Effect of Social Security on Personal Savings*, Cambridge, Mass., 1974.

———, *The Future of Social Security*, Washington 1976.

J. Pencavel, "Some Labor Market Implications of the Payroll Tax for Unemployment and Old Age Insurance," mimeo., Stanford Univ. 1974.

J. Smith, "On the Labor-Supply Effects of Age-Related Income Maintenance Programs," *J. Hum. Resources*, Winter 1975, *10*, 25–43.

R. Weiss and B. Schiller, "The Value of Defined Benefit Pension Plans: A Test of the Equalizing Differences Hypothesis," mimeo., Univ. Maryland 1976.

City of Los Angeles vs. Manhart, No. 76-1810, U.S. 55 L. Ed. 2d., 1978.

U.S. Department of Labor, "Questions and Issues Relating to the Proposed Amendments to the Age Discrimination in Employment Act of 1967," mimeo., 1977

The Market for Ph.D. Economists: The Academic Sector

By CHARLES E. SCOTT*

The market for Ph.D. economists has changed drastically in the last twenty years. During the 1960's there was a "boom" market with the number of Ph.D.s granted tripling, the salaries being received by new Ph.D.s rising by a third and the number of entering graduate students in economics rising by 2.6 times its 1960 level. Since the early 1970's there has been a decline in all of these variables. Other graduate fields experienced booms at roughly the same time as economics but the boom in economics has not been followed by a "bust" of the proportions seen in other fields (for example, see Richard Freeman, 1971). Nonetheless there has been a change in the nature of placement in the economics market. It is more difficult for candidates of given qualifications to obtain a desirable job, and the probability of being unemployed or underemployed at the time of receipt of the degree has increased. Many individuals who would have qualified for academic jobs at highly ranked institutions are now accepting nonacademic jobs or academic jobs at lower ranked institutions.

In order to analyze these changes in the market for economists a two-pronged approach has been applied to the academic portion of the market. Although the analysis applies specifically to the academic portion of the market, the conclusions reached are more general, since the majority of economists work in academia. The discussion is divided into three major sections. The first presents the results of a cobweb analysis of the market, recognizing the inherent feedbacks in the market while ignoring quality variations in both jobs and applicants. The second section addresses the variations in quality. The third gives conclusions which can be drawn from the combination of the two approaches.

I. Aggregate Model

Economics occupies an intermediate position among doctorate disciplines, characterized by relatively diverse sources of demand and an "average" 1960's experience with respect to changes in salaries, enrollments, and numbers of Ph.D.s granted. The lack of continued increase in entering enrollments since the 1967–68 academic year, which led to the decline in the number of Ph.D.s granted, has combined with the diverse demand to lead to relatively good current market conditions. In order to analyze the interrelationships and the lags in the market and to predict the future potential, a cobweb model was estimated, similar to those applied by Freeman to other disciplines. The diversity of the demand in this market and the corresponding lessened variance of salaries and degrees created problems but a reasonably good fit was obtained using this model.

The estimated model recognizes the economic responsiveness of potential entrants into graduate school to currently available information about salary offers to new Ph.D.s and financial aid available for graduate studies in economics. The Ph.D. equation recognizes explicitly the link between the number of degrees awarded and enrollments a few years earlier as well as to changes in the economic incentive to finishing the degree. The demand equation describes the relation between the number of Ph.D. economists

*Assistant professor, Marquette University. The research for this paper was supported by a doctoral dissertation grant from the Manpower Administration, U.S. Department of Labor. Points of view or opinions in this document do not necessarily represent the official position or policy of the Department of Labor. A longer version of this paper is available upon request.

demanded, and academic salaries and demand proxies. The model is closed with a market-clearing equation and a stock generation equation. The model is specified in first *log* difference form in order to recognize the proportionality of the relationships (as opposed to linearity) and to limit the effects of the inherent multicollinearity in the data.

Salary (real) is the equilibrating mechanism during the market period since supply is predetermined by past supply decisions. As will be seen in Section II this is an oversimplification of what actually happens, but it is necessary to allow estimation of the model. Thus combining the demand and market-clearing equations and making salary the dependent variable gives the final, estimatiable, equation. The estimates of the coefficients of the three equations are given in Table 1.

The *log*-linear specification leads to interpretation of the coefficients as proportional relations and the constant term as a trend. The Ph.D. equation shows no significant trend in the number of Ph.D.s. The changes in the number of Ph.D.s are, instead, dependent on changes in the number of first-year full-time graduate students in economics with an elasticity of one-half. In addition, changes in the relative salaries being received by new economics Ph.D.s in academia cause a more than proportional change in the number of new Ph.D.s awarded four academic years later. Both independent variables are serving as proxies for a distribution of these same variables which would affect decisions. The enrollment lag corresponds to the average registered time for completion of the Ph.D. degree in economics in recent years. The (real) salary term corresponds to the salaries being offered new graduates during the second year of graduate school when the decision between completion and stopping short of the Ph.D. with a masters degree would be contemplated most seriously.

The first-year enrollment equation shows a significant trend in the first-year full-time enrollments in economics of 5.2 percent. This

TABLE 1 COBWEB MODEL OF THE MARKET FOR PH.D. ECONOMISTS

Equations	R^2	F
1965-74 $\Delta ln\ PhD = 020 + 4936^{b}\Delta ln\ FFT_{t-6} + 1\ 25^{b}\ \Delta ln\ (SAL/Prof)_{t-4}$ (.023) (.180) (469)	.798	13 79[a]
1961-74 $\Delta ln\ FFT = .052^{b} + .887^{c}\ \Delta ln\ SAT_{t-1} + 164^{c}\ \Delta ln\ Sup_{t} - 160^{b}\ DRF_{t}$ (017) (411) (087) (.061)	.736	9 30[a]
1964-74 $\Delta ln\ SAL = .027 - .061\ \Delta ln\ Stock_{t-1} + 228^{a}\ \Delta ln\ FFT_{t-2} + .223^{a}\ \Delta ln\ R\&D_{t-1}$ (050) (914) (074) (.075)	812	12.98[a]

Sources: PhD = Ph.D.s awarded in economics (U S. Office of Education, *USOE*).

FFT = First-year full-time graduate student enrollment in economics (*USOE*)

SAL = Starting salary of academic economists in constant 1967 dollars (Francis Boddy data linked with *AEA* salary data).

R&D = Federal research and development spending in economics (*NSF, Federal Funds . .*).

Prof = Median earnings of male professionals, technical, and kindred workers in constant 1967 dollars (U.S. Bureau of the Census).

Sup = The number of federal fellowships and traineeships awarded (Freeman, 1974; *NSF, Graduate Science Education, . . .*).

DRF = 1 in 1967 and 0 otherwise. Draft deferments for graduate school were dropped Fall 1968.

Stock = Stock of Ph.D. economists based on an assumed retirement and death rate of 1.8 percent (Cartter, 1971), the number of Ph.D. recipients in economics who accepted domestic placement (*NRC, Summary Report*) and an estimate of the stock of economics Ph.D.s of 9,678 as of Fall 1972 (*NSF*, 1973).

[a]Significant at the 1.0 percent level.

[b]Significant at the 2.5 percent level.

[c]Significant at the 5 percent level.

may be due to the increase in the size of the age cohort who would consider graduate school and have the necessary qualifications for entry or credentials inflation. Attempts to account for the former effect proved unsuccessful and did not appreciably affect the constant term. In addition to the trend, changes in the starting salaries being received by new economics Ph.D.s are proportionally reflected in enrollment changes. Changes in graduate student funding also affect the value of the degree and consequently enrollments. The level and probability of funding is the desired measure, but, lacking this, a general proxy—the number of federal fellowships and traineeships—was used. The final variable, a dummy, accounts for the elimination of draft deferments in 1968 which had previously served as an artificial incentive to attend graduate school.

In the salary equation, the strong correlation between the supply variable and the *R&D* variable led to the former's coefficient being nonsignificant. Estimates of the equation without the *R&D* term yielded a significant negative supply term but omission of the *R&D* demand variable leaves out an important influence. The high correlation among the various academic demand measures requires inclusion of only the first-year full-time enrollment in graduate school in economics, which enters with a lag reflecting the resistance to staffing changes in response to what might be temporary changes in demand. *R&D* funding enters with a lag due to funding making available new openings which are subsequently filled.

Using these three equations and the stock generation equation, projections were generated and are given in Table 2. They indicate that the annual decline in real salaries of 5 percent which began in 1973 will slow to a decline of approximately 1.6 percent into the 1980's. This will lead to a continuation of the decline in the number of degrees awarded, but will not deter entry into graduate school. Enrollments are actually projected to increase even with there being no change in the availability of student aid. If there is a decline in funding, this trend may be moderated or reversed. This projected increase is not too unreasonable with the need for further education becoming more important.

TABLE 2—MARKET INDICATORS IN THE MARKET FOR PH.D. ECONOMISTS

Year	First-year Full-time Enrollment in Economics	Ph.D.s in Economics	Starting Salaries in Economics[c]
1970–71	3,320[a]	787[a]	$10,472[a]
1975–76	3,127[a]	872[a]	8,856[a]
1980–81	3,558[b]	763[b]	8,114[b]
1985–86	4,225[b]	711[b]	7,504[b]

[a]Actual figures.
[b]These projections are based on the assumptions of a 14 percent foreign placement of new Ph.D.s (the average for academic 1957–58 through 1975–76); 1.8 percent retirement and death rate for Ph.D economists (the estimate which Cartter used), a 2.6 percent growth rate for federal research and development spending in economics (*NSF*'s projections for totals), a 2 percent growth in professional earnings, and no change in graduate student support.
[c]Nine-ten month basis, 1967 constant dollars.

II. Placement

So far the discussion has ignored the quality and time dimensions, treating Ph.D.s as homogeneous and the market period as a point in time. These are simplifications necessary for aggregate analysis but they ignore the actual market workings. Therefore explicit reference to quality of applicant, quality of academic placement, and timing of placement were needed to more fully describe market behavior.

A. *Quality*

The method of addressing quality is that suggested by Lester Thurow. He contends that individuals compete for jobs based on their relative costs of being trained for the job rather than on the basis of salary. Under this hypothesis the market is seen as two queues—prospective employees and job vacancies. In the market for Ph.D. economists, degree recipients compete on the basis of background characteristics with the market clearing through alteration of the quality of applicant who fills a given caliber vacancy.

The assumption here is that the background characteristics—the primary one being the rank of the graduate school from

TABLE 3—MARKET CONDITION AND PLACEMENT INDICATORS

Year	Applicant/ Vacancy Ratio[a]	PD Index	Salary[b]	SEEK
1961	.94		8,116	4.9
1962	.81		8,278	3.7
1963	.65		8,942	4.7
1964	.68		9,419	4.1
1965	.81		9,788	3.2
1966	.83		10,288	3.5
1967	.75	−1.43	10,125	1.8
1968	.77	−1.52	10,797	3.1
1969	.97	−1.58	10,929	3.4
1970	1.24	−1.48	10,472	4.6
1971	1.48	−1.83	9,969	6.2
1972	1.42	−1.65	10,176	6.0
1973	1.17	−1.67	9,454	6.4
1974	1.15		9,230	6.8
1975	1.13		8,856	7.2

Sources: Applicant/Vacancy Ratio, *AEA*, unpublished data; *PD* Ratios, calculated by author, Salary, Boddy, *SEEK*, *NRC*, unpublished doctorate records employment plans data.

[a]Weights for the moving average are 1 for t − 1, 2 for t, and 1 for t + 1. This was done to minimize randomness.

[b]Nine–ten month basis, 1967 constant dollars

which the degree was received—indicate the candidate's potential as a researcher and teacher. Since placement is based on perceptions of quality rather than actual quality, the graduate school rank is a good first "screen" for ordering the applicant queue for analysis and conforms to market behavior.

First-employment placement affects not only the immediate working conditions—teaching load, caliber of coworkers and students, availability of research funding, etc.—but also future potential income through enhancing or detracting from the accumulation of good work experience. Since the employer queue is ordered by applicants' perceptions, and the characteristics making a graduate school well respected will make it a desirable professional environment, the same ranking scheme can be used for the employers' queue as for the applicants' queue. This common measure of quality used here is the Roose-Anderson ranking of graduate programs with the addition of a fifth group for nonranked Ph.D. granting institutions and a sixth for institutions not granting Ph.D.s.

The first hypothesis tested was whether placement worsened during periods of decreasing demand and improved during periods of increasing demand. The measure of quality of placement (called a Push Down (*PD*) index to signify the forcing of the graduates of the lower ranked schools to placement in the lesser ranked schools):

$$(1)\quad PD = \frac{\sum_{i=1}^{5}\sum_{j=1}^{6} a_{i,j}^{(r_i - r_j)}}{\sum_{i=1}^{5}\sum_{j=1}^{6} a_{i,j}}$$

where $a_{i,j}$ is the number of graduates of the schools in group i who obtained their first job in schools of rank j; r is the number of the group (i.e., rank of the schools); i is the subscript applying to the school of graduation; j is the subscript applying to the school of first placement. Although simple, this measure indicates the average change in rank from graduate school to first placement and assumes that the further down in the employers' queue an individual is placed, the worse the quality of placement. In general, *PD* will be negative indicating "downstream" placement of graduates of a school in lesser ranked schools.

A moving average of the number of applicants per vacancy registered at the annual American Economic Association's (*AEA*) job placement service was used as a measure of market condition. This measure was tested against the *PD* index for a negative relationship, with more applicants per vacancy causing worse placement. The Spearman rank correlation was significantly negative (−.714) at the 5 percent confidence level, supporting the hypothesis. Thus, placement is acting as an equilibrating mechanism.

B. *Timing*

Applicants and employers have an alternative to changing their expectations and/or requirements. That is to delay finalizing an employment agreement, hoping for a better match (job or applicant) to materialize. During "suppliers" markets applicants are likely to fulfill their expectations and have no incentive to delay, and employers will have no incentive to await a downward revision of the

applicants' expectations. The converse is true during "demanders" markets and hence timing of placement should be later during these years.

Since placement, rather than the existence of jobs, is the potential problem, a "snapshot" of the percentage of Ph.D.s who are still seeking employment as of receipt of their degree would measure this timing effect. This variable (*SEEK*) turns out to be significantly negatively correlated with the *PD* index (−.857, Spearman rank correlation) supporting variation of placement timing's being another of the variables in the market-clearing process.

C. *Effect on Salary*

The variation of timing and placement according to market condition suggests that there may be some carry-over from year to year. Thus after a year when quality of placement is low and timing of placement is later, market expectations will be lowered, including salary expectations. In order to test this feedback relation both the *PD* index and the *SEEK* variable were tested for correlation with the following year's salary. The *PD* index proved not to be correlated with the next year salary term suggesting that quality of placement serves as an alternative to salary adjustment rather than a prelude. The *SEEK* variable, however, was significantly correlated with the following year's salary (−.618). This indicates the interyear effect may occur by the later placement causing lower salaries the following year, as employers react, with a lag, to the changed supply-demand balance.

III. Conclusion

Combination of the alternative analytical approaches shows that consideration of placement or salaries alone gives only a partial picture. The decline in salaries is actually an underestimate of the loss in value of the Ph.D. in economics. Quality of placement, which fell initially, appears to have rebounded partially suggesting a slowed market decline. Timing of placement appears to be occurring later each year, foretelling a future of poor prospects. The sum of these is a bleak picture of lower salaries with placement lower than during the late 1960's.

A major unaccounted for factor, which can only worsen the prospects in economics, is the general decline in higher education projected for the 1980's. With salaries already projected to fall, and the number of degrees granted not projected to drop appreciably, the rebound of placement—our bright spot—is lessened in credibility. This leads to the expectation of a dismal market in which to practice our dismal science.

REFERENCES

Francis Boddy, "Survey of Academic Starting Salaries for Economics Departments," unpublished, various years.

Allan Cartter, "Whither the Market for Academic Economists," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 305–10.

———, *Ph.D.'s and the Academic Labor Market*, New York 1976.

Richard Freeman, *The Market for College Trained Manpower*, Cambridge 1971.

——— **and David Breneman,** *Forecasting the Ph.D. Labor Market: Pitfalls for Policy*, Washington 1974.

Kenneth D. Roose and Charles J. Anderson, *A Rating of Graduate Programs*, Washington 1970.

Lester Thurow, *Generating Inequality*, New York 1975.

American Economic Association (*AEA*), unpublished annual placement service applicant and vacancies data.

———, unpublished Universal Academic Questionnaire starting salary data.

———, *Directory of Members of the American Economic Association*, Chicago 1974.

National Science Foundation (*NSF*), *Characteristics of Doctoral Scientists and Engineers in the United States*, Washington 1973.

———, *Federal Funds for Research and Development*, Washington, various issues.

———, *Graduate Science Education, Student Support and Post-doctorals*, Washington, various issues.

U.S. Bureau of Census, *Statistical Abstract of the United States*, Washington, various

issues.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington 1978.

U.S. National Research Council (*NRC*), *Summary Report, Doctorate Recipients from United States Universities*, Washington, various issues.

———, unpublished doctorate records employment plans data tabulated for and supplied to me by Richard Freeman.

U.S. Office of Education (*USOE*), *Earned Degrees Conferred*, Washington, various issues.

———, *Students Enrolled for Advanced Degrees*, Washington, various issues.

Stocks and Flows of Academic Economists

By BARBARA B. REAGAN*

Hiring and promotion of faculty comprise the heart of current efforts to obviate discrimination in universities. Policy changes can be observed in flows long before any sizable effect can be detected in the stock data. This paper considers three questions related to the overall question of whether employment opportunities in universities in the economics profession are opening to women: 1) What is the current stock of women economists in academia, with emphasis on women with new Ph.D.s in economics; 2) are the flows of women economists into the various faculty levels (a) in line with the proportion of women in the relevant stock, and (b) enough greater than the past pattern to suggest affirmative action is occurring; and 3) can the revolving door syndrome be quantified, and is it affecting women disproportionately?

These questions pick up the problem at the point of production of Ph.D.s. No analysis is made of the prior issues of reducing barriers to filling the pipe line with women earning Ph.D.s, or of the contributing issues related to encouragement of women after employment in academia.

Administrators in universities want to know about stocks of women economists and the hiring pool characteristics so as to estimate their chances of effecting affirmative action. Those interested in public policy and evaluation of recent affirmative action policies in universities need to assess the flows in employment in relation to the appropriate stocks. As Barbara Bergmann has noted, the personnel actions involved in the flows are the points at which any taste for discrimination is implemented. Here is where any monopsony power which is used in ways disadvantaging women is brought to bear, and role bias coming from the culture may be perpetuated by decision makers.

In this paper the numbers of faculty hired, the numbers released or who quit, and the numbers promoted and/or given tenure are traced. It is assumed that the degree-granting institutions assure that the pools of women and men receiving Ph.D. degrees are of equal quality.

Under a hypothesis of random hiring, a binomial distribution is used to compute the probability of observing a number greater than or equal to the actual number of women hired or promoted, given the proportion of women in the stock and the number of persons selected from the stock (i.e., the flow). This approach is used for the summary of personnel actions for all economics departments for 1977–78, 1976–77, and 1974–75. Because the number of personnel actions is relatively low when the departments are subdivided, a more simplistic method is used by type of department. Obviously, the proportions of men (P_m) are readily attainable from the proportions of women (P_w) because $1 - P_w = P_m$. If the proportion of women hired is five percentage points less than the proportion in the relevant stock, the proportion of men among those hired will be five percentage points greater than the proportion of men in the stock.

To summarize the results, evidence of affirmative action in economics departments is sparse over the last four years. The entry door at the assistant professor level is still narrow for women, and the revolving door syndrome is working too well for many women economists at the assistant professor level. Any efforts on behalf of women at the filter points have been so slight that the sex composition of faculties has changed little. The applicant pool of new Ph.D.s is running 8–11 percent women. Those women who are there are undoubtedly highly motivated. Improved opportunities will be needed to avoid discouragement of new Ph.D. candidates.

*Professor of economics, Southern Methodist University. Additional data and analysis are available from the author. Comments by Francine Blau, Ann Friedlaender, and Myra Strober, and special tabulations and computations by Charles Scott and Patricia Kirby Cantrell are appreciated.

I. Data Base

The data for this analysis are taken from the reports made by economics departments in colleges and universities to the American Economic Association (*AEA*), starting in 1974–75, on the Universal Academic Questionnaire (*UAQ*). (For the 1976–77 and 1977–78 reports, departments in Canada were also included.) Data by sex for 1974–75, 1975–76, and 1976-77 have been published on an unweighted basis (see the author, 1976, 1978). The 1977–78 data have not been available previously. The 1975–76 data on a weighted basis are not available for this paper.

Data are presented by four departmental groupings: Departments that offer Ph.D. degrees in economics and belong to the Chairmen's Group, the major producers of Ph.D. degrees; all other Ph.D.-granting departments of economics; departments with the M.A. the highest degree offered in economics; and those with the B.A. the highest degree offered in economics. Data for all departments are weighted averages of the grouped data.[1] This procedure holds constant variation in the sample size by type of department.

II. Stock of Women Economists in Academia

From 1956 to 1970 the proportion of women among Ph.D. recipients averaged a little over 4 percent and from 1971 to 1974, it was 6 percent (see William Spellman and D. Bruce Gabriel). Data reported on the *UAQ* showed that in 1973–74 and 1975–76 the upward trend continued (8 and 11 percent). In 1976–77 the proportion of women awarded Ph.D.s in economics dropped below 9 percent at the same time that the total number of Ph.D.s awarded dropped 12 percent. The number of Ph.D.s awarded women in 1976–77 was about 9 percent less than in 1973–74. In 1976–77, the number of M.A. degrees in economics fell only 5 percent, and the number of B.A. degrees in economics increased about 5 percent. Among both M.A. and B.A. degree recipients, the number of women increased 20 percent. Thus the potential supply of women who might be interested in seeking Ph.D.s in economics is higher than it was in 1973–74.

The drop in number of women receiving Ph.D.s in economics in 1976–77 was not accompanied by a disproportionate drop in financial aid that year or the following year Thus the decrease in women may be recouped subsequently.

Women receiving Ph.D.s in economics are distributed among the top Ph.D. granting institutions in the same proportions as are men (see Spelling and Gabriel). The major field of specialty selected by women economists differs somewhat from that of men. Nearly 20 percent of the women have labor as the major specialty, compared with 10 percent for the profession as a whole. Consumer economics is also female intensive. Conversely, the male intensive areas are business, finance and marketing, and agriculture and natural resources. Among younger women increasing numbers are now going into monetary and fiscal economics (see Myra Strober and the author, pp. 309–11). Differences in fields of specialization may disadvantage some women economists in hiring decisions. It should not affect promotion and tenure decisions at the same school or release decisions because the department presumably would not hire a person with a given set of specialties unless needed.

The recent slump in the job market for faculty is well recognized as college enrollments decelerated and the ratio of new Ph.D.s to enrolled students has increased (see Rich-

[1] Estimates based on the average number per school such as data on B.A. degrees granted and students enrolled are weighted by the relative number of departments in each group in the universe (.06, 08, .08, and .78, respectively). Data on the percentage of women among faculty stocks and flows are weighted by the relative number of full-time faculty in each type of department (.20, .23, .12, and .45, respectively). The data for assigning these weights come from *AEA* records, the number of departments awarding degrees in economics from the *U.S.* Office of Education, Department of Health, Education and Welfare, the average number of faculty per department by type of department from the *UAQ*, and the number of departments awarding M.A. degrees as the highest degrees in economics as a percent of departments awarding graduate degrees from the *UAQ*. The proportions of departments reporting on the *UAQ* in 1977–78 were more than 70 percent of the Chairmen's Group and of the other Ph.D. departments, more than 50 percent of the M.A. departments, and more than 25 percent of the B.A. departments.

and Freeman, p. 32). Although it was more difficult for academic economists to find a job in the mid-1970's than it was in the 1960's, overall the number of full-time professors, associate professors, and assistant professors in all economics departments reporting on the *LAQ* increased 8–10 percent annually in the years studied. Shifts in the composition of academic offerings favored economics over other disciplines.

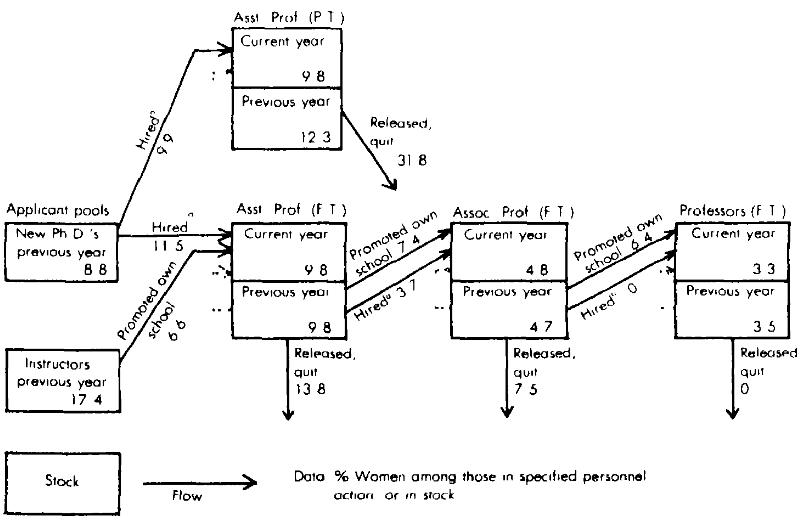
Although the subsequent analysis will show that the pattern of affirmative action (or the converse) is far more complex, increases in women faculty tend to occur only when total faculty increases. For example, in 1977–78, a 7–14 percent increase in all full-time economists at professorial ranks was accompanied by an 11–29 percent increase in the total number of women in these ranks in each of the four departmental groups.

III. Job Turnover of Academic Economists

Labor turnovers (see Figure 1) provide opportunities for affirmative action. If, in addition, the number of faculty needed is expanding as has been true in economics, there is even more room for opening employment to women holding the appropriate scholarly credentials.

To give an overall appraisal for all departments of economics in each of the three years studied, the personnel action at each of the filter points is considered under a hypothesis of random selection. A binomial distribution is used to compute the probability of observing a number greater than or equal to the actual number of women hired or promoted, given the proportion of women in the stock and the number of persons in the personnel action. The hypothesis of random hiring is rejected when this probability is less than 7 percent.

In 1974–75 and again in 1977–78, the hypothesis of random hiring is rejected for women hired at the assistant professor level, and thus the operation of affirmative action is suggested; for example, for 1977–78, 11.5 percent of those hired were women with 8.8 percent women in the stock of new Ph.D.s, a difference of 2.7 points. The probability that a number greater than or equal to this would occur under random hiring is only .06. Similarly, the proportion of women assistant professors promoted in their own schools to associate professor in 1976–77 and the proportion of associate professors hired as full professors in other schools in 1974–75 and 1976–77 were greater than could be expected under random promotion and hiring. This



[a]Includes some hired in lateral transfer at same rank

FIGURE 1. JOB TURNOVER OF ACADEMIC ECONOMICS, ALL DEPARTMENTS, 1977–78

suggests affirmative action at these points in the specified earlier years, but not in 1977–78. Then in 1977–78, the revolving door syndrome reversed the affirmative action for assistant professors. The number of women at the assistant professor level who were released was greater than could be expected under random release; for example, for 1977–78, 13.8 percent of the assistant professors released were women compared with 9.8 percent women in the stock of assistant professors, a difference of four points and a probability of a number greater than or equal to this under random hiring equal to .05.

One other nonrandom selection occurred. In 1974–75 the number of women among the assistant professors who were promoted to associate professors was so low (3.6 percent compared with 10.1 percent women in the pool of assistant professors) that the probability of a number less than or equal to this with random promotion was less than .02. This and the selections of women at most of the other filter points in each of the three years that are consistent with a hypothesis of random selection suggests no affirmative action effect for women in economics departments except at the few points noted.

To appraise affirmative action by type of department, the evidence on all ten filter points from hiring assistant professors up through hiring or promoting to full professors is summed into an index (−10 to +10). A plus impact is counted if the flow proportion is more than two percentage points greater than the relevant stock proportion. In the index a plus impact at one point counteracts a negative impact for women at another point. Only the economics departments in which the B.A. is the highest degree offered had a positive score (3 in 1974–75 and 2 in 1976–77). No group of departments had a positive affirmative action score for women in 1977–78.

Discrimination against women at the associate professor level in hiring by other schools occurred in 1977–78 in Ph.D. and M.A. departments, when a greater than or equal to five percentage point difference in the proportions unfavorable to retention of women is used as an indicator of discrimination. In the Chairmen's Group, sex discrimination is indicated at this flow point and in promotions to associate professor in all three of the pairs of years observed. Although the pattern by type of department varies, the incidence of discrimination was about as frequent by this measure in 1977–78 as in 1974–75.

The action occurring at the entry level of assistant professors may well be of the greatest interest. The affirmative action at that hiring point, and the subsequent revolving door action which more than wipes out the small gain, condition the possibilities for affirmative action at higher levels and warrants further analysis by type of department.

Four indicators of affirmative action at the assistant professor level are 1) the proportion of women hired as assistant professor is more than two percentage points higher than the proportion of women in the new Ph.D. pool of applicants; 2) the proportion of women released is not more than two points greater than the proportion hired; 3) the proportion of women assistant professors increased more than two points from the previous year; and/or 4) the year-to-year increase in the number of women assistant professors employed is more than five points greater than the increase in all assistant professors. In the years considered, as many as three of the four affirmative action indicators for assistant professors were met only twice—other Ph.D. departments in 1977–78, and B.A. departments of economics in 1976–77. Mathematically, affirmative action is possible in all schools (i.e., all types of schools could hire greater than two percentage points higher than the proportion of women in the stock) because only about half of the new Ph.D.s, female or male, go into academia.

The revolving door syndrome for those women assistant professors hired was much in evidence in the years studied. The indicators used are that the proportion of women assistant professors 1) released is greater than or equal to the proportion hired, 2) promoted in their own schools is greater than two points below the proportion in the stock the previous year, and 3) in the current year stock is greater than two points less than in the previous year. All three of these indicators of a revolving door were seen in each of the four types of departments in 1974–75. Two or more were seen in each type of department in 1977–78. All three were observed in Ph.D.

departments in the Chairmen's Group throughout the period. The evidence suggests that the revolving door syndrome was slightly less in 1977–78 than it was in 1974–75 for women assistant professors in departments other than those in the Chairmen's Group, but no group of departments is free of all the indicators.

At the upper professorial ranks, the indicators of affirmative action are infrequent. In considering flows at the more senior ranks, a crucial question is whether the previous flow at the assistant professor level over a longer run has been sufficient to provide a base of women ready for consideration for promotion and/or tenure. In general, this is true. The proportions of women hired as assistant professors in all departments were 12, 9, and 12 percent in 1974–75, 1976–77, and 1977–78.

Tenure decisions are also vital aspects of the flow process, particularly at the associate professor level. A small matched sample of departments reporting on the *UAQ* in both of the last two years suggests that the proportion of women among those awarded tenure was down in 1977–78 from 1976–77.

In 1977–78, 61 percent of the departments of economics still had all-male faculties. The best record in the years studied, that is, the lowest proportion of all-male economics departments, tends to be in the Chairmen's Group (25–30 percent), and the worst record is in the B.A. departments (64–70 percent all male) The small matched sample for the last two years suggests no change in these relative positions. The continued existence of all-male departments is an area where opportunities for women economists at any level need to be opened as fast as the supply of women Ph.D.s permits if we hope to educate well women and men for the future. By example, all-male departments discourage the flow of women into the profession's pipe line and exemplify most clearly the need for affirmative action. In other departments the disproportionate revolving door syndrome for women at the assistant professor level needs to be slowed, and opportunities for women economists at the other ranks improved.

In conclusion, the stock-flow measure suggested in this paper provides a first approximation to a workable indicator of discrimination. It is a more elemental approach than regression techniques, and avoids some of the complications that have led to misuse of regression results. It can suggest areas where micro-data analysis is needed. The simplicity of the stock-flow method permits use of alternative assumptions as to the speed of affirmative action thought to be reasonable. It permits quantification of the revolving door syndrome, and highlights the current state of affirmative action at the filter points well before the results (or lack of results) show up in the stock data.

REFERENCES

B. R. Bergmann, "Data Needs Relating to Fighting Employment Discrimination Against Women," in Barbara B. Reagan, ed., *Issues in Federal Statistical Needs Relating to Women*, forthcoming.

R. B. Freeman, "The Job Market for College Faculty," disc. paper no. 596, Harvard Instit. Econ. Res., Harvard Univ., Dec. 1977.

B. B. Reagan, "Report of the Committee on the Status of Women in the Economics Profession," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 490–501; May 1976, *66*, 509–20; and May 1978, *68*, 484–99.

W. E. Spellman and D. B. Gabriel, "Graduate Students in Economics, 1940–74," *Amer. Econ. Rev.*, Mar. 1978, *68*, 182–87.

M. H. Strober and B. B. Reagan, "Sex Differences in Economists' Fields of Specialization," in Martha Blaxall and Barbara B. Reagan, eds., *Women and the Workplace, The Implications of Occupational Segregation*, Chicago 1976.

U.S. Office of Education, *Earned Degrees Conferred: 1969–70 Institutional Data*, Washington 1970.

Mobility in the Labor Market for Academic Economists

By DAVID E. AULT, GILBERT L. RUTMAN, AND THOMAS STEVENSON*

Until recently, most studies of the market for academic economists did not explicitly consider the effects of differences in the set of skills possessed by each individual in the market on his mobility. Howard Tuckman, James Gapinski, and Robert Hagemann, and William Becker departed from this pattern by characterizing the market for economists as a market for specific skills. Their studies analyzed observed variations in faculty salaries associated with differences in acquired teaching, research, and service skills. The individual decision to invest in each of these skills is separated from the decision to enter a specific discipline. Once the individual has acquired the minimum skill levels necessary to enter the academic market, he/she acquires additional units of each skill in response to the reward structure of the institutions at which he/she is or desires to be employed.

This paper offers a new approach to the study of the movement of economists among academic institutions, based on the idea of acquired skill differences. The total supply in each academic market is composed of individuals with different quantities and qualities of skills in teaching, research, and service. The number of combinations of skills available is limited only by the number of individuals in that market. At one extreme are faculty members with large investments in research skills and minimal investments in teaching and service skills. At another extreme are those individuals whose investments are largely concentrated in teaching; and at the third extreme are those who have concentrated their investments in service skills.

In order for a market for academic skills to exist, there must be a demand for these skills. Academic institutions determine their demands for skills by estimating the quality and quantity of teaching, research, and service skills required to achieve a set of institutional goals. These desired skill levels and quantities are then used to screen applicants for faculty openings. When academic units such as economics departments enter the market to fill a faculty opening, they must measure, no matter how crudely, the potential contribution of each candidate to the output of the department.

The purpose of this study is to examine the movement of economists within academe Because research skills are the most easily measured and, perhaps, the most marketable of the three skills sought by academic institutions, this study will analyze the effect of differences in acquired research skills on the movement of economists among departments. Because adequate measures of teaching and service skills have yet to be developed, no attempt will be made to analyze the effects of differences in acquired teaching and service skills on mobility. A theoretical explanation of the demand for economists by academic institutions will be presented. The flow of research skills among departments will then be analyzed.

I

In this study, economics departments are treated as production units that use a production function to determine the mix of inputs required to produce an output or set of outputs. Departments, similar to any output-producing unit, attempt therefore to maximize their outputs as defined by the mission and goals of the institution in which they are housed. Even though the output of a department may be difficult to define and quantify, it is a function of the teaching, research, and service skills embodied in its

*Professors and associate professor of economics at Southern Illinois University-Edwardsville and St Louis University, respectively. We wish to thank Dennis Schlott and William Waymire, who were responsible for transforming the raw data. Any errors that remain are our sole responsibility.

faculty and the complementary inputs available to the faculty.

In particular one can hypothesize that the output of department i in year t, Q_{it}, is maximized subject to a production function and a budget constraint:

$$(1)\quad Q_{it} = f(T_{i1t}, \ldots, T_{int}, R_{i1t}, \ldots, R_{int}, S_{i1t}, \ldots, S_{int}, SS_{it}, K_{it})$$

and

$$(2)\quad B_{it} = \sum_{n} P_{ijt}X_{ijt} + P_{it}^{SS}SS_{it} + P_{it}^{K}K_{it}$$

where

T_{ijt} = the teaching skills possessed by faculty j of department i in year t

R_{ijt} = the research skills possessed by the faculty j of department i in year t

S_{ijt} = the service skills possessed by faculty j of department i in year t

SS_{it} = the support of services provided by the institution in which department i is housed in year t

K_{it} = the physical capital stock to which the faculty of department i has access in year t

B_{it} = the operating budget of department i in year t

P_{ijt} = the salary paid to faculty j in department i during year t

X_{ijt} = faculty j who is a member of department i during year t and who possesses skills T_{ijt}, R_{ijt}, and S_{ijt}; $j = 1, \ldots, n$, the number of faculty in department i

P_{it}^{k} = the price paid by department i for each input k in year t.

The amount invested by each individual to acquire additional units of teaching, research, and service skills depends upon the benefits relative to the costs of such investments. The principal cost to the individual is the value of the time required to produce an additional unit of each skill. The returns from such investments are determined by the effects of these investments on the marginal products of these skills. These marginal products, which are the primary determinants of faculty compensation, are directly related to the skill levels possessed prior to admission into a graduate program and acquired in graduate school. The former consist of general knowledge, study habits, and discipline—related knowledge gained primarily from a baccalaureate education. Upon acquiring the minimum skills necessary to enter the academic market, the returns to individual investments in teaching, research, and service skills depend upon the skill levels of other faculty as well as the quantity and quality of complementary inputs such as computing facilities, library holdings, internal and external research support, and the presence of other universities or research institutes in the area. At time t, the value of an individual to each potential academic employer is a function of his marginal product in each skill area.

Although the supply of economists may be relatively elastic over time in any given year t, the supply of skills is fixed and equal to the sum of the skills possessed by each of the economists who have met the minimum entrance requirements, although only a fraction of these economists are actively seeking a new position in t. The demand for academic economists in year t is equal to the sum of all academic positions although the number of openings, which determines the number who can move, is less than this total. If the institutions that produce the highest quality output employ those faculty with the highest skill levels, these institutions should offer the most desired combinations of salary and other forms of compensation because, in a competitive market, compensation will reflect marginal product. Several problems are encountered when attempting to test this hypothesis with existing salary data. Compensation includes not only the present salary but expected future salary, fringe benefits, and other sources of income. An individual may accept a lower salary in period t in order to increase his expected future earnings through investment to acquire additional skill units. The present value of the expected income stream over the planning horizon m for individual j is, therefore, equal to

$$(3)\qquad W_{j0} = \int_0^m E(W_t)_j e^{-rt}\, dt$$

where

W_{j0} = the present value of the income stream of economist j at time 0

$E(W_t)_j$ = the expected salary of j in year t

r = the discount rate

t = the planning horizon of length m of j.

$$(4) \quad E(W_t)_j = F(W_o, A_t, M_t, OJT_i)$$

where

W_0 = current nine-month salary of j

A_t = j's access to the editorial staffs of major journals during m

M_t = j's access to the program committees of major economics meetings

OJT = on-the-job training accruing to j as a result of professional experience in i.[1]

Given that there is a market for academic skills and that the development of research skills is a continuous process: 1) departmental demand for research skills will depend upon the output mix produced by the department; 2) the perceived quality of the department's output will vary directly with the level of research skill demanded; 3) individuals who invest in order to acquire additional research skills expect to be compensated for such investments; 4) because research skills are a more general form of human capital than teaching or service skills, they should be the most marketable of the three skills; and 5) individuals who successfully invest in additional research skills should therefore flow to those departments that produce the highest quality outputs and offer the greatest compensation for research skills. The complex nature of the output of an economics department, however, complicates empirical investigation of the academic market for economists. The product mix as well as the quality of each product in the mix vary from department to department. Although most economics departments treat research as a complement to teaching skills, departments that offer a Ph.D. program are expected to demand more and higher levels of research skills than departments that offer only baccalaureate degrees in economics. There may be at least two markets for research skills: one in which research skills command a positive price and a second market in which the price for these skills is zero.

[1]Teaching and service skills are usually acquired through on-the-job experience while research skills are acquired through graduate education as well as experience.

II

Several testable hypotheses concerning the movement of economists are generated from the theoretical analysis presented. To test the hypothesis that graduate training in economics is dependent upon the quality of general training received, the following estimating equation was used:

$$(5) \quad PHD_j = K + \beta_1 BA_j + \sum_{\gamma=1}^{n} \beta_{1+\gamma} PREF_{\gamma j} + u \quad \beta_1 > 0$$

where

PHD_j = the quality rating of the institution from which j received his doctorate

BA_j = the quality rating of the institution from which j received his baccalaureate

$PREF_{\gamma j}$ = a vector of dummy variables indicating the regional and employment preferences of j.[2]

The manner in which the academic market allocates economists to faculty openings was examined using the following models:

$$(6) \quad FJ_j = K + \beta_1 PHD_j + \beta_2 YRSPHD_j + \beta_3 INST_j + \sum_{\gamma=1}^{n} \beta_{3+\gamma} PREF_{\gamma j} + u$$
$$\beta_1 > 0; \beta_2 < 0; \beta_3 < 0$$

[2]*REGION* = 1 if the region of the institution from which j received his Ph.D. was the same as the region in which the institution that granted the baccalaureate was located; *REG*, *REGm*, and *AREAm* were the regional preference variables in *FJ* and *MOVE* models.

SAME = 1 if the institution from which j received the baccalaureate and Ph.D. were the same; *SM* = 1 if *FJ* and Ph.D. institutions were the same.

PRIVATE = 1 if the baccalaureate and Ph.D. granting institutions were private; *PRV*, *PRPRm*, *PRPBm*, *PBPRm*, and *PBPBm* indicated the type of institution that j moved to and that he left in the other models, *PBPBm* was the excluded category.

HOMEm = 1 if the institution to which j moved and the institution from which j received his baccalaureate were located in the same region. *BAHM* = 1 if the institutions were the same.

INSTRm, *ASSTm*, *ASSOCm*, and *PROFm* indicated the rank held by j prior to move m; *ASSTm* was the excluded category.

where

FJ_j = the quality of the institution at which j was employed after completion of his Ph.D. requirements

$YRSPHD_j$ = the number of years between the year in which j received his baccalaureate and the year in which his Ph.D. was received

$INST_j$ = a dummy variable = 1 if j was employed at the same institution, g, both prior to and immediately after completion of his Ph.D. requirements.

$$(7)\quad MOVE_{mj} = K + \beta_1 sJ_j + \beta_2 YRSsJ_j + \beta_3 PUBIND_j + \beta_4 YLPROM_j + \beta_5 PROM_j + \beta_6 ADMIN_j + \sum_{\gamma=1}^{n} \beta_{6+\gamma} PREF_{\gamma j} + u$$

$$\beta_1 \gtrless 0;\ \beta_2 < 0;\ \beta_3 > 0;\ \beta_4 < 0;\ \beta_5 < 0;\ \beta_6 < 0$$

where

$MOVE_{mj}$ = the difference in the quality of institutions in which j was employed, $m = 1, \ldots, z$; for move 1 = (rating of institution 2 − rating of institution 1)

sJ_j = the quality of the institution s at which j was employed prior to move m

$YRSsJ_j$ = the number of years that j spent at institution s

$PUBIND_j$ = the publication index of j through time t − 1 where t is the year in which move m was made

$PROM_j$ = a dummy variable = 1 if j received a promotion with move m

$ADMIN_j$ = a dummy variable = 1 if j received an administration appointment with move m

$YLPROM_j$ = the number of years since the last promotion of j prior to move m.

With respect to rating economics departments, the Roose-Anderson (R-A) rating of economics departments that offered Ph.D. programs was used.[3] *PUBIND* is a weighted index of j's journal publications and was constructed as follows:

$$(8)\quad PUBIND_j = \sum_{q=0}^{t-1} \sum_{v=1}^{15} w_{vq} JRN_{vqj}$$

where

JRN_{vqj} = the number of publications in journal v during year q by j; $v = 1, \ldots, 15$; and $q = o$, the year in which j accepted his first position, . . . , t − 1

w_{vq} = an index number derived from Carol McDonough indicating the quality of journal v relative to the other journals in the sample.

The variables in the $PREF_{\gamma j}$ vectors in each model were included to capture the influence of locational and institutional preferences, if any, on the quality of Ph.D. school, department of first employment, and departments of subsequent employment. Geographic preferences or the desire to return to or be near specific institutions may be sufficient reason for j to move from one institution to another.

III

The results of the *OLS* analysis for Ph.D. School, First Job after Ph.D., and *MOVE* models are summarized in Table 1. Although one individual in the sample moved eight times during the period, only the first and second moves generated a sufficient number of observations for analysis. The results offer limited support for the hypothesis that there is a market for research skills because the range of movement due to acquisition of additional research skills is small.

The quality of undergraduate and graduate education received is a major determinant of an academic economist's career pattern. The quality of undergraduate training is one of the principal determinants of the quality of grad-

[3]Four other indices were constructed to rate the quality of each university in the sample. The results using these indices support the results reported.

TABLE 1—RESULTS[a]

Ph.D School		First Job		First Move		Second Move	
K	2.55 (0.06)[d]	K	0.85 (0.085)[d]	K	1.81 (0.16)[d]	K	1.82 (0.43)[d]
BA	0.41 (0.02)[d]	PHD	0.44 (0.023)[d]	1J	−0.69 (0.044)[d]	2J	−0.67 (0.096)[d]
REGION	−0.27 (0.09)[d]	YRSPHD	−0.0021 (0.0019)	YRS1J	0.017 (0.021)	YRS2J	0.025 (0.032)
PRIVATE	0.38 (0.05)[d]	INST	−0.11 (0.058)[b]	PUBIND	0.0034 (0.0014)[c]	PUBIND	0.0053 (0.0028)[b]
SAME	−0.40 (0.095)[d]	REG	0.018 (0.087)	YLPROM	−0.011 (0.023)	YLPROM	−0.0041 (0.036)
R^2	.7748	SM	0.78 (0.096)[d]	PROM	−0.42 (0.12)[d]	PROM	−0.58 (0.26)[c]
S.E. F-Ratio n	0.91 161.77[d] 169	PRV	0.70 (0.061)[d]	ADMIN	0.21 (0.21)	ADMIN	0.35 (0.29)
		PRPB	0.97 (0.091)[d]	REG1	0.050 (0.16)	REG2	−0.084 (0.33)
		PRPR	0.048 (0.064)	AREA1	0.44 ([illegible])[d]	AREA2	0.69 (0.31)[c]
		R^2	.1825	INSTR1	0.023 (0.11)	INSTR2	−0.14 (0.70)
		S.E. F Ratio n	0.94 145.56[d] 1868	ASSOC1	0.016 (0.14)	ASSOC2	0.081 (0.22)
				PROF1	0.089 (0.25)	PROF2	0.60 (0.37)
				PBPR1	0.045 ([illegible])	PBPR2	0.31 (0.18)
				PRPB1	0.76 (0.13)[d]	PRPB2	−0.077 (0.24)
				PRPR1	[illegible] (0.14)[d]	PRPR2	0.42 (0.27)
				BAHM1	0.071 (0.25)	BAHM2	0.11 (0.59)
				HOME1	−0.14 (0.19)	HOME2	0.22 (0.66)
				R^2	.5116 (.4926)	R^2	.4816 (.3894)
				S.E. F Ratio n	[illegible] [illegible][d] [illegible]	S.E. F-Ratio n	0.88 5.2251[d] 107

[a]Standard errors are in parentheses

[b]Significantly different from zero with a probability of .10.

[c]Significantly different from zero with a probability of.05.

[d]Significantly different from zero with a probability of .01.

uate training received. A segmented academic market based upon the perceived quality of the undergraduate and graduate education received by the individual seems to exist, with the quality of formal education setting an upper bound to future mobility. In fact, the positive signs and significance of the parameter estimate for *SM* support the results obtained in earlier studies that individuals who obtained their first academic position at the same institution from which they earned their doctorates are employed by higher quality institutions than those at which new Ph.D.s not retained are employed. Further, all other things equal, the highest rated institution at which an academic economist will be employed is the institution of first employment. Each move tends to be to a lower rated institution. Those who are attracted to a second or third position through the offer of a promotion or an administrative post accept employment at institutions with lower ratings than the institutions at which they were first employed. It appears, therefore, that lower rated institutions use promotions and administrative positions as a means of attracting faculty from more prestigious institutions. Departments producing the higher quality output demand new faculty with a higher level of demonstrated research skills than lower rated departments. The *ABD*'s seem to be employed by lower rated departments than individuals who have completed all Ph.D. requirements.

The strongest evidence that there is a market, though perhaps segmented, for skills is provided by the signs and significance of the parameter estimates for *PUBIND*. Given that publications are evidence of research quality, the more publications by an individual, and the more prestigious the journals in which they appear, the higher is the institutional rating of the individual's second or third position. However, publications appear to have less influence on the quality of the third job obtained than the second. This supports the hypothesis that those with successful investments in research skills flow to institutions that emphasize graduate education and reward research output—the higher rated schools. Upon accepting an initial academic position, publications appear to be the principal means by which an academic economist can move to a more prestigious department. While upward mobility is possible, the parameter estimates indicate that a considerable investment in research skills on the individual's part is required. Large jumps, even for those who invest successfully, are unlikely.

The longer an individual is at institution *s*, the more specific capital in the form of teaching and service skills he tends to accumulate and the greater should be his value to that institution relative to other potential employers. Most specific capital is concentrated in teaching and service. Years in job *s* and the *YLPROM* parameter estimates are, however, insignificant. The *YLPROM* estimates may be insignificant because individu-

als not promoted at the end of their probationary periods are typically denied tenure and are forced to move to a lower quality school where they receive their first promotion. The *PREF* vectors attempt to isolate the influence of regional biases on the part of the employing institutions and economists seeking employment. The results indicate that regional biases play a role in screening economists for academic openings as well as in individual acceptance of an academic position.

Because the nature of the data permits analysis only of the effects of individual differences in research skills on mobility, the results cannot be conclusive with respect to the thesis that the academic market is a market for skills. A complete test requires data on the individual's evaluation of the compensation streams that are expected to result from alternative life cycle investment patterns and the effects of search costs on academic mobility. The results do provide support for the hypothesis that a market for research skills does exist, although it may be highly segmented. The quality of baccalaureate and graduate education is a primary determinant of the range of institutions that will consider employing an individual. The quantity and quality of an economist's publications do improve his upward mobility but the increase in upward mobility is slight. More research remains to be done to determine the effects of institutional and individual geographic and other preferences on mobility in order to test the hypothesis that the academic market for economists is segmented.

REFERENCES

W. E. Becker, Jr. "The University Professor as a Utility Maximizer and Producer of Learning, Research, and Income," *J. Hum. Resources*, Winter 1975, *10*, 107–15.

C. McDonough, "The Relative Quality of Economics Journals Revisited," *Quart. Rev. Econ. Bus.*, Spring 1975, *15*, 91–7.

Kenneth Roose and Charles Anderson, *A Rating of Graduate Programs*, Washington 1970.

H. Tuckman, J. Gapinski, and R. Hagemann, "Faculty Skills and the Salary Structure in Academe: A Market Perspective," *Amer. Econ. Rev.*, Sept. 1977, *67*, 692–702.

The Cartelization of World Commodity Markets

By ROBERT S. PINDYCK*

The cartelization of world commodity markets is not a new phenomenon. In a historical survey of the experience of some international commodity cartels, Paul L. Eckbo shows that at one time or another there has been an attempt to cartelize the market for most of the major internationally traded commodities. The large majority of these attempts at cartelization, however, were failures—the cartel either dissolved after a short period of time, or in some cases the cartel remained in force officially, but had little or no real impact on price and member revenues. Of the fifty-one formal cartel organizations documented by Eckbo, only nineteen could be considered successful in the sense of being able to maintain a price significantly higher than what it would have been in the absence of agreements. But even the successful cartels were limited in their durability; the average lifetime of the formal agreements was about five years, and only five of the nineteen cartels lasted ten years or longer.

What is new is the growing concern that the prospects for *successful* cartelization have suddenly become greater, and that in the future, world commodity markets are likely to be increasingly dominated by cartels. Much of this concern, of course, has been the result of the Organization of Petroleum Exporting Countries' (*OPEC*) spectacular success in quadrupling world oil prices, and the International Bauxite Association's (*IBA*) success in tripling the price of bauxite. Warranted or not, this concern now casts a shadow over predictions, policy prescriptions, and proposals for the international "management" of commodity markets. Buffer stocks and other instruments for price stabilization, for example, become the vehicles for cartelization and the establishment and maintenance of the monopoly price. And for some, cartelization, or more specifically, an implicit or explicit transfer of monopoly and monopsony power from developed to developing countries, is an essential and justifiable component of the New International Economic Order (*NIEO*).[1]

Given the historical success record of international cartels, is there any reason to expect new attempts at cartelization to succeed where similar attempts in the past have failed? Has the structure of world commodity markets—or the environment surrounding them—changed in such a way as to better facilitate the formation and success of cartels, so that over the next decade we are likely to witness a proliferation of international cartels that will succeed in raising the prices of a large number of key commodities?

There are no simple answers to these questions. While the *interest* in cartelization on the part of some *LDC*s may indeed be greater, there appears to be no clear change in the structure of commodity markets that would facilitate their cartelization. It is difficult to agree with C. Fred Bergsten's assertion, for example, that the environment has shifted to one in which supplies of raw material commodities are shrinking as demand keeps growing, thereby encouraging cartelization.

In the past some cartels succeeded while others failed for reasons specific to each market and to each cartel configuration. As a more recent example, *IBA* succeeded while *CIPEC*, the copper cartel, did not—and

*Massachusetts Institute of Technology. The research on which this paper is based was funded by the National Science Foundation, under grant #DAR78-19044.

[1]This argument is made very nicely by Carlos Diaz-Alejandro, who has termed the *LDC*'s spearheading the drive for the *NIEO* the "new oligopolists."

cannot—succeed because of some major differences in the structures of the bauxite and copper markets. Evaluating the prospects for successful cartelization in the future will have to be done on a market-by-market basis, in each case determining whether a proposed cartel configuration actually has the potential to raise and maintain a monopoly price.

I. The Requisites for Successful Cartelization

There are really two issues involved in the potential for cartel success. The first has to do with problems of cartel organization and stability. The success of a cartel depends on its members coming to an agreement on price and production levels, and on the division of revenues. Also required is a means of detecting and deterring cheating on the part of cartel members. These organizational problems have been the focus of most attempts to evaluate the future prospects for cartelization. The likelihood that existing and potential cartels can solve these problems has often been used as the major argument supporting forecasts of a proliferation of cartels. See, for example, Bergston and Zuhayr Mikdashi.

A cartel may indeed solve its organizational problems, but this does not guarantee its success. The existence of monopoly profits depends not only on market concentration, but also on the elasticity of demand. A cartel may account for 100 percent of the production of a commodity, but if the demand curve facing the cartel is highly elastic, there will be little room to raise price, and little prospect for obtaining significant monopoly profits.

The potential for monopoly profits, assuming that the organizational problems can be solved, is the second issue involved in cartel success. In fact, it may be the most important issue; if monopoly and competitive prices differ only slightly, no organizational strategy can result in cartel success. On the other hand, the ability to solve the organizational problems may depend very much on the potential for monopoly profits. There are costs involved in the organization and maintenance of a cartel—costs associated with the determination of price and the rationalization of output and revenues, political costs that may result when output is reduced or revenues are used to finance stockpiles, and costs associated with the risk of being undercut by other cartel members that might try to make large short-term profits outside the cartel boundaries. Cartel members must be willing to bear these costs if the cartel is to succeed, and doing so will be so much easier the greater the potential monopoly profits.

To evaluate the prospects for an increase in the cartelization of world commodity markets we must therefore evaluate the potential gains that cartelization could bring. Calculating these gains, however, may not be straightforward; a simple comparison of the monopoly and competitive prices in static equilibrium is likely to be misleading, since much of the gain from cartelization may depend on the dynamics of the market.

II. Measuring the Potential Gains from Cartelization

The structure of most commodity markets is inherently dynamic. In many markets, for example, the demand for the commodity will respond to price only with long lags, as will the supply from countries that produce competitively. A cartel itself may face long lags in increasing or decreasing its production capacity. A major increase in production by a coffee cartel, for example, could occur only after the four or five years needed for new trees to reach maturity, and *decreases* in production might involve even longer lags, since governments might be able, through incentives, to restrict more tree planting, but might not be able to destroy existing trees. And for many mineral resources, depletion problems may critically affect intertemporal pricing and production decisions.

Because of the dynamic structure of commodity markets, the comparison of short-run or long-run static monopoly and competitive profits is simply not a useful way of measuring the potential gains from cartelization. Large short-run profits might be available to a cartel which has little potential for long-run gains, while resource exhaustion, through its tendency to bring competitive and monopoly prices closer together, might eliminate what a static analysis would indicate are

large potential gains. The gains from cartelization can be better measured by assessing the *equity value* of the production activity under different market structures—that is by comparing the present discounted value of the flow of current and future profits under cartelization (beginning at the moment the cartel forms) to that under a competitive market, assuming in both cases that production descisions are made optimally.

These comparisons can in fact be made by constructing fairly simple models of the commodity markets in question, and determining the cartel price trajectories and sum of discounted profits using an appropriate optimal control algorithm. I have done this for three commodities—oil, bauxite, and copper—measuring the potential gain as the ratio of the sum of discounted profits under cartelization to that under competition. (See my 1977, 1978 papers.) The models provided robust estimates of the gains and their distribution over time, and in each case also yielded surprisingly accurate *ex post* forecasts of cartel pricing behavior. The results indicated that the gains from cartelization are considerable for oil and bauxite, but negligible for copper. The *CIPEC*'s inability to enjoy large monopoly profits arises only in part from its smaller market share (35 percent), but also from the dynamic response of *secondary* competitive supply (from scrap), where, because of the stock effect, the short-run elasticity is much larger than the long-run one.

Stylized models such as these should also be applicable to the analysis of other commodity markets. Such a model would begin with a dynamic function relating total demand for the commodity to prices and income. This function should capture the important time lags, and, for a commodity such as bauxite where demand is extremely inelastic up to a limit price at which point it becomes extremely elastic, may be highly non-linear. Next, a function for competitive (noncartel) supply would describe the dynamic dependence of that variable on prices. For an exhaustible resource, the supply function should include the effects of depletion. For commodities such as copper and other metals, it might consist of two or more equations that would explain both primary and secondary supplies. Cartel demand is then total demand minus competitive supply. The model is closed with a cartel cost function, which might depend on reserve levels if the commodity were a depletable resource, or a commodity such as coffee or timber where the planting of trees is involved might depend on current capacity, which in turn would be a function of lagged prices and past capacity.

The cartel price trajectory and sum of discounted profits could then be found as the optimal control solution for such a model. The determination of the price trajectory and sum of discounted profits in the competitive case would depend on the characteristic of the commodity in question. For an exhaustible resource, the price is such that rents grow at the rate of interest, and exhaustion occurs just as the demand for the commodity becomes zero. For a produced good such as coffee, the cartel members act competitively and produce at the point where price is equal to marginal cost. In this case, market clearing yields the price at each point in time.

III. Cartel Organization and Stability

Given that the potential gains are present, a cartel can succeed if solutions can be found to the problems of organization and stability. The question, then, is whether a group of producers (with a combined market share that could yield sufficient potential monopoly gains) can agree on an optimal aggregate production level (thereby determining the cartel's contract surface), agree on a division of output and profits (i.e., select a point on the cartel contract surface—a sharing rule), and find a means to detect and deter cheating. As D. K. Osborne has shown, the most critical of these problems are the determination of the contract surface and the detection of cheating; once solutions to these problems have been found, solutions to the sharing problem and the deterence problem follow readily.

What is important here is that for many real or imagined cartels there are no inherent obstacles to solving the problems of finding the contract surface and detecting cheating. As a result, for many commodities there may be several cartel configurations that could be

reasonably stable, so that the real problem, once again, may be the existence of monopoly gains.

There may, of course, be hurdles that must be overcome in solving the organizational problems. As Bjarke Fog pointed out, determination of the contract surface may be complicated by differences in opinion over the characteristics of demand and competitive supply, the existence of substitutes, etc. More important, cartel members may operate under different constraints (for example, reserve levels in the case of mineral resources) and/or have different time preferences. However, as shown by Esteban Hnyilicza and myself for a model of *OPEC*, these differences become simple differences in objectives (for example, different discount rates in the calculation of the equity value of the production activity), and solutions exist for the reconciliation of these differing objectives, just as they do for the sharing problem in general.

As for the detection of cheating, this need not be a very serious problem, particularly if the cartel produces a good that is fairly homogeneous (for example, oil) or if sales can to some degree be centralized. Much of the success of the iodine cartel, the longest lived in Eckbo's survey (1878–1939), can be attributed to the fact that all sales were made through a central cartel office in London. Of course, consuming countries might—and should—take actions to encourage cheating by making its detection more difficult.[2] Unfortunately, rather than adopting measures to exacerbate cartels' organization and stability problems, consuming countries are now considering policies for commodity market intervention that would aid cartels in solving these problems. Any measure that has the effect of centralizing sales or disseminating information on transactions prices works to help solve the detection problem—which is why buffer stocks, marketing arrangements, and national or international purchasing agencies could work so much to the disadvantage of the consuming countries.

IV. The Prospects for Commodity Cartelization

It seems to me that over the next decade or two the success of international commodity cartels will depend more on the existence of potential monopoly gains than on the ability to solve the organization and stability problems. As explained above, there is no inherent reason why the organizational problems cannot be solved, and the consuming countries, through various policies of market intervention, may even assist in the solution. If the potential monopoly gains do exist, that will provide all the more motivation to solve the other problems.

For many commodities, however, the potential monopoly gains may simply not exist. As argued above, this is why for practical purposes *CIPEC* is a cartel on paper only. And with the exception of bauxite, cartels in most other mineral markets would probably have the same problems that *CIPEC* has had.[3] Cartels in markets for such minerals as iron ore, manganese ore, lead, tin, zinc, and nickel would face demands that become relatively elastic in the long run, competitive secondary supplies (from scrap) with large *short-run* price elasticities, as well as income elasticities of demand which, according to Benison Varon and Kenji Takeuchi, are well below one for many of these metals. As for nonmineral commodities, such as coffee, cocoa, timber, and pulp, it is difficult to assess just how large the potential gains might be without actually constructing a market model such as that described earlier. On the one hand, there are considerable lags involved in increasing noncartel production capacity, so that a cartel of, say, the major Latin America coffee producers might be able to make significant short-run profits. On the other hand, the demands for most of these commod-

[2]For example, by adopting import ticket plans such as the one proposed by Morris Adelman for crude oil. Under the Adelman plan, tickets which give the holders the right to import oil are sold in an anonymous auction, and can be freely transferred or resold. The system would encourage cheating by permitting *OPEC* countries to establish brokers who would bid for and purchase tickets in the United States, thereby discounting the price of their oil in exchange for assured sales.

[3]Bauxite is unusual in that its world demand is extremely inelastic (up to a limit price), even in the long run (see the author, 1977). Another resource market where cartelization appears to be successful is uranium

ities respond much more quickly to price than is the case for many of the mineral resources.

The measurement of the potential gains from cartelization will require careful analysis on a market-by-market basis. The indications are, however, that there are not many markets where these gains are likely to be large. Undoubtedly more cartel organizations will form, but we should not expect many of them to have a significant impact on commodity markets.

REFERENCES

M. A. Adelman, "Oil Import Quota Options," *Challenge*, Jan./Feb. 1976, *18*, 17–22.

C. F. Bergsten, "The Threat Is Real," *Foreign Policy*, Spring 1974, No. 14, 84–90.

C. F. Diaz-Alejandro, "International Markets for LDC's—The Old and the New," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 264–69.

P. L. Eckbo, "OPEC and the Experience of Some Non-Petroleum International Cartels," M.I.T. Energy Lab. work. paper, June 1975.

B. Fog, "How Are Cartel Prices Determined?," *J. Indus. Econ.*, Nov. 1976, *5*, 16–23.

E. Hnyilicza and R. S. Pindyck, "Pricing Policies for a Two-Part Exhaustible Resource Cartel: The Case of OPEC," *Eur. Econ. Rev.*, Aug. 1976, *8*, 139–54.

Z. Mikdashi, "Collusion Could Work," *Foreign Policy*, Spring 1974, No. 14, 57–68.

D. K. Osborne, "Cartel Problems," *Amer. Econ. Rev.*, Dec. 1976, *66*, 835–44.

R. S. Pindyck, "Cartel Pricing and the Structure of the World Bauxite Market," *Bell J. Econ.*, Autumn 1977, *8*, 343–60.

———, "Gains to Producers from the Cartelization of Exhaustible Resources," *Rev. Econ. Statist.*, May 1978, *60*, 238–51.

B. Varon and K. Takeuchi, "Developing Countries and Non-Fuel Minerals," *Foreign Aff.*, Apr. 1974, *52*, 497–510.

National and International Policies Toward Food Security and Price Stabilization

By DAVID BIGMAN AND SHLOMO REUTLINGER*

Developing countries are justifiably concerned about year-to-year variations in food grain production, inside and outside their boundaries. For large segments of the population who live at the margin of adequate nutrition, short food supply and high food prices mean curtailment of consumption to unacceptably low levels. High food prices lead to upward pressure on wages and have other undesirable macro-economic consequences. Low food prices can erode farmers' incomes and adversely affect future production. High international prices may cause serious balance-of-payments problems for food importing countries.

To cope with the problems of instability within the framework of the market system, a country can use a range of policy instruments that may be divided into three categories: operating buffer stocks, adjusting foreign trade, and implementing price subsidy and support programs for specific groups and sectors. However, each policy or combination of policies while having a desirable effect on one objective may have an undesirable effect on other objectives. In addition, some programs cannot be implemented effectively without drawing on resources beyond the means of most developing countries.

In this paper we analyze the effectiveness of intervention policies in the national and the international food grain markets in achieving prespecified stabilization goals. The study is based on simulation experiments with a model of a developing economy, with parameters chosen to approximate orders of magnitude of a country like India.

*International Monetary Fund and World Bank, respectively. The views presented are our own and are not to be attributed to these institutions. Research on which this paper is based is part of a study undertaken by the World Bank (RPO 671-24) on various aspects of food stabilization in developing countries.

I

The model is of an open economy engaged in trade with the rest of the world. The sources of fluctuations in the domestic market are assumed to be random disturbances in domestic supply and in the international price. The country is assumed to be "self-sufficient" in the sense that in a "normal" year, when both the country's production and the world price are at their mean level, there would be no price differential between the country and the world and thus no incentive for trade. In other years, however, uncorrelated random fluctuations in domestic supply or in the world price would tend to create price gaps which would trigger imports or exports in a free market.

The main components of the model are:[1]

(i) A world grain price model which estimates the world price on the basis of (randomly distributed) world production and an estimated world demand function.

(ii) A country grain market model which estimates grain consumption, exports and imports, inventories in stock, and domestic grain price for each level of country production and world market price, in accordance with the policies implemented by the government.

(iii) A system of decision rules which represent the government policies in the grain market.

(iv) A procedure for the calculation of gains or losses to the various consumer groups, to producers, and to the government, due to stock, price, and trade policies.

Estimates are generated for a large sample of grain productions and aggregated into frequency distributions.

[1]The detailed report describing the models and the parameters is available upon request from the authors.

Consumers in the country are divided into three groups: low-income urban consumers, medium- and high-income urban consumers, and rural consumers. A separate demand function is specified for each group. The model can accommodate the specification of a food price subsidy program for low-income urban consumers and a floor price policy for farmers. Stabilization by means of buffer stocks can also be incorporated; it is specified as holding supply within predetermined boundaries and within the limits of storage capacity constraints.

Trade between the country and the world is assumed to be determined by market forces within the limits of specific trade policies implemented by the government. The instruments for enforcing government goals with respect to trade are tariffs. The model can accommodate a wide range of trade policies, including a policy aimed at increasing food supply and price stability in the country. Such stabilizing policy could be implemented by the government paying a subsidy to importers for each imported ton of grain or by the government importing grain and selling it in the market at a lower price in times of domestic production shortfall or high international prices.[2] Similarly, the government could pay a subsidy on each ton of exported grain to prevent a sharp decline of prices to producers, when domestic production is high or international prices are low.

II

For many governments, the primary objective of intervening in grain markets is to ensure a regular flow of supplies to consumers and to meet the needs of vulnerable sections of the population. The effects of interventions on the long-run welfare of the economy and the distribution of national income, which we discuss in the next section, are only secondary considerations for most governments. Any expected economic losses to certain groups or sectors, and possibly also to the economy at large due to the stabilization policies, might then be regarded best as an *insurance premium* paid by the market participants against the risks of scarcity and famine.

The most noteworthy result of our simulation experiments presented in Table 1 is the highly stabilizing effect of international trade. Contrary to a rather widely held belief, free trade increases rather than diminishes the stability in the domestic market. Isolation from the world market, aimed at avoiding the external "shocks" to prices and supplies associated with trade, will actually result in higher rather than lower instability. Free trade can reduce the probability of an extreme shortfall in the quantities available to low-income consumers and the fluctuations in prices to an extent that no reasonably sized buffer stock can achieve. It should be noted, however, that these results are due to the particular choice of parameters in our model. International trade will be less stabilizing (and possibly even destabilizing) the more stable is the domestic supply and the less stable is the foreign supply. Trade will also be less stabilizing the higher is the price elasticity of demand and the country's share in the world trade of grain. Elsewhere, however, we have demonstrated that for most of, if not all, the developing countries, trade is likely to have a strong stabilizing effect.

Against the option of isolating the domestic market from the world market, which increases rather than decreases instability, a stabilizing trade policy insulates the country from external disturbances with compensating subsidies and taxes, thereby increasing the flow of trade in and out of the country. The effects of this policy on the stability of the domestic market are indeed far-reaching as illustrated in Table 1. While it is clear that food security for consumers and stability of farmers' income can be achieved by various combinations of trade and buffer stock policies, Table 1 shows that these stabilizing effects cannot be achieved without destabilizing the balance of payments and the government's budget. Trade which acts to stabilize prices and supplies results in unstable foreign exchange transactions. The internal price policies (subsidy program and support price),

[2] The former policy is essentially the one implemented in Indonesia, whereas the latter is closely related to the trade operations of the Indian government.

TABLE 1—STABILITY OF FOOD POLICY RELATED GOALS: PROBABILITY OF SPECIFIED EVENT (Shown in Percent)

	Food Security Targets			Financial Accounts	
Policies	Grain Consumption of Low-Income Urban Consumers Below 95% of Median	Grain Price Above 120% of Median Price	Farmers' Income Below 90% of Median	Fiscal Costs of Food Policies in Excess of $500 Million	Food Import Bill in Excess of $500 Million
Closed Economy					
No Intervention	30.4	23.6	28.9	–	–
6MMT Stocks	19.2	14.3	15.3	7.8	–
12MMT Stocks	14.5	10.8	9.9	11.2	–
Internal Price Programs	6.2	32.5	26.6	51.3	–
Internal Price Programs + 6MMT Stocks	3.6	20.9	13.0	34.0	–
Open Economy					
Free Trade	22.1	11.8	16.5	–	11.2
Free Trade + 6MMT Stocks	12.3	6.2	9.9	6.8	9.4
Stabilizing Trade	3.9	0.6	4.1	0.2	22.4
Stabilizing Trade + 6MMT Stocks	2.2	0.3	2.6	5.3	14.3
Free Trade + Internal Price Programs	0	16.4	15.4	30.1	16.2
Stabilizing Trade + Internal Price Programs	0	5.9	3.6	14.2	22.4

while very effective in reducing instability for the beneficiary groups, substantially increase the instability for other groups or sectors.

Table 1 also illustrates how the stabilizing effect of any particular policy depends on other prevailing forms of intervention. For instance, buffer stocks will significantly increase the stability both of food grain prices and farmers' income in a closed economy, but will have only small stabilization benefits in an open economy. On the other hand, when the internal subsidy programs are in effect, stocks will be particularly effective in reducing the instability of cost related parameters, such as the food import bill and the government budget.

III

Welfare gains from the various stabilization policies are measured by the expected change in producer's and consumer's surplus. The model estimates both the incremental gains from a policy, given the policies which are already practiced, and the total gains from a given combination of policies.

Table 2 illustrates the expected incremental gains and losses from a 6MMT buffer stock under alternative sets of existing policies. The more open the economy, the less frequently are buffer stocks used to stabilize supply. Thus, the expected gains or losses to the various groups and sectors from the storage operation (as well as its stabilizing effect) are significantly smaller in the open economy.

The expected gains or losses from stocks among some groups can vary significantly depending upon what other policies are already practiced. It should be emphasized that these gains are the marginal gains from stocks. Thus, for example, producers' losses from stocks when the internal price program prevails are in effect only a reduction in their gains from the subsidy program.

An intriguing result is the effect of storage on the government's budget and on the balance of trade. When the government is implementing subsidy and support price programs, buffer stocks yield large savings in government expenditures on these programs

TABLE 2—EXPECTED ANNUAL GAINS AND LOSSES DUE TO A 6MMT BUFFER STOCK
(Millions of Dollars)

Policy	Economy	Consumers: Low Income	Consumers: Other	Producers	Government: Storage Operation	Government: Other	Government: Total
Without Internal Price Subsidy and Support Program							
No Trade	−43	−32	−56	104	−59	0	−59
Free Trade	−69	8	25	−25	−76	−1	−77
Stabilizing Trade	−62	−9	−20	31	−81	17	−64
With Internal Price Subsidy and Support Programs							
No Trade	−34	−100	85	−310	−51	342	291
Free Trade	−72	−36	32	−73	−79	84	5
Stabilizing trade	−67	−14	2	0	−82	27	−55

which far outweigh the cost of operating the buffer stock. These significant savings on the government's account could justify buffer stocks in spite of "negative" effects on low-income consumers and on producers, particularly when the losses incurred by these groups from stocks consist merely of reductions in their gains from other policies. Buffer stocks may also have some justification on account of their stabilizing effects on foreign exchange requirement.

Table 3 summarizes the expected gains and losses from the internal price subsidy program. Although implementation of the program involves a very small net loss to the economy as a whole—and is therefore possibly a price worth paying for securing a minimally adequate level of nutrition for all—the transfer of income involved is quite substantial. The essence of the program, in effect, is a redistribution of the quantity of grain available for consumption via the price system. By subsidizing the price of grain to low-income consumers, the government forces other consumers to pay a higher price. The end result will therefore be a transfer of food grain from high- and middle-income consumers to low-income consumers.

IV

We have seen that an active trading policy can be an effective, least costly way of resolving food security problems. However, such a policy implies a highly unstable food import bill. In years of stress, countries would continue to be confronted with having to make a bitter choice between disrupting their development efforts or getting along with less than adequate food supplies. International arrangements involving both stocks held in or on behalf of developing countries and a financial standby facility to assist countries to cope with unusually high food import bills, would be, therefore, consistent with the broad objectives of international development assistance.

To calculate the size of a food import bill

TABLE 3—EXPECTED ANNUAL GAINS AND LOSSES OF PROGRAM TO SUBSIDIZE CONSUMPTION OF LOW-INCOME CONSUMERS
(Millions of Dollars)

Policy	Economy	Consumers: Low Income	Consumers: Other	Producers	Government: Storage Operation	Government: Other	Government: Total
No Trade	−25	301	−617	880	0	−589	−589
Free Trade	−9	244	−170	244	0	−327	−327

TABLE 4—PAYMENTS BY A FOOD IMPORT BILL INSURANCE SCHEME TO ELIMINATE CONSUMPTION SHORTFALLS IN DEVELOPING COUNTRIES

Payments Needed to Secure Food Grain Consumption at Average Level	Probability of Payments (%)	
	Without Buffer Stock	With 10 MMT Stock
Below $2 Billion	33.4	23.1
Above $2 Billion	13.9	7.6
Expected Annual Payments ($ Million)	742	435
Expected Cost of Storage ($ Million)		50

insurance scheme which would protect all developing countries' food grain supplies from falling below the average level, we used a slightly modified version of the previously discussed simulation model. Specifically, we assumed that the average food grain import gap of the developing countries is 30 MMT, their domestic food grain supply is distributed normally with a mean of 300 MMT and a standard deviation of 11 MMT, and the import price of food grains is distributed normally with a mean of $150 per ton and a standard deviation of $30 per ton.

Table 4 indicates the size of an insurance fund needed to secure food grain supplies of the developing countries at their average supply level. Payments from the fund provide for any excess in the food import bill beyond its normal level. Even without stocks, the expected payments would not be very high. A buffer stock which is operated to counter fluctuations in both domestic supply and import price would substantially reduce the "premium" cost of attaining food security. (See also P. Konandreas, B. Huddleston, and V. Ramangkura and Reutlinger.)

V

Food grain production in today's world proceeds under a great diversity of ecological and technological conditions. Poor harvests in some regions are offset by favorable harvests in other regions. Internal and international trade have been greatly facilitated by sophisticated means of transportation and needy areas can draw on the supplies available in other areas. At current food production levels, supplies are adequate to feed the world's population even in lean years. A basic premise of our study has been that current food crises are characterized not by overall scarcity but by gross maldistribution of food. With the existing distribution of income and wealth among individuals and nations, the free market economy is unable to prevent frivolous uses of food at the same time that other people are starving. While trade and buffer stocks can play a major role in stabilizing food grain supplies, effective insurance against hunger will also require special financial measures. Even poor nations can seek to safeguard against their vulnerability to climatic instability by intervening in the market to secure minimally adequate consumption for all, by means of price discrimination or rationing in favor of low-income consumers. Our analysis shows, however, that such national food security measures are very costly, involve substantial transfers of income, and require massive intervention by the government in the free market with possible undesirable consequences in the long run. Alternatively, or at least in addition, we propose an international financial undertaking which would enable developing countries to acquire food in times of need.

REFERENCES

D. Bigman and S. Reutlinger, "Food Price and Supply Stabilization: National Buffer Stocks and Trade Policies," unpublished paper, May 1978.

D. E. Hathaway, "Grain Stocks and Economic Stability: A Policy Perspective," in *Analysis of Grain Reserves*, U.S. Department of Agriculture, ERS-634, Aug. 1976, 1–11.

P. Konandreas, B. Huddleston and V. Ramangkura, "Food Security: An Insurance Approach," res. rept. no. 4, Int. Food Pol. Res. Inst., Sept. 1978.

S. Reutlinger, "Food Insecurity: Magnitude and Remedies," *World Develop.*, 1978, *6*, 797–811.

Measuring the Impact of Primary Commodity Fluctuations on Economic Development: Coffee and Brazil

By F. GERARD ADAMS, JERE R. BEHRMAN, AND ROMUALDO A. ROLDAN*

That fluctuations in primary commodity export markets have significant impacts on the economies of the producing countries is intuitively clear and widely accepted. But there is little consensus on the direction and significance of these impacts on economic development. Fluctuations in the value of primary commodity production and exports have the potential for affecting the growth path through a number of channels. This paper is concerned with delineating these channels of influence and with establishing quantitative dimensions for their effects. As an illustration we focus on coffee in relation to the economy of Brazil. The relationships between the coffee producing sector and the national economy are established econometrically and integrated into a macro model of the Brazilian economy. Alternative scenarios for coffee production and exports illustrate the nature of the impacts throughout the economy and the potential effects on Brazilian growth.

I. Structural Approach to Commodity Markets: Macro-Economy Interactions

The numerous empirical studies of the correlation between export market performance and economic growth reach equivocal conclusions. The cross-section approaches obscure the ways in which the economy of the producing country is affected. What, for example, are the forward and backward linkages? What are the impacts on fiscal and external balance? What is the effect on income distribution?

A number of econometric model studies have recognized the role of export producing sectors in the context of developing country models. C. Rangarajan and V. Sundararajan's study, which takes a time-series modeling approach using small models for several countries, obtained inconclusive results and argued that "... the usefulness of international schemes for stabilizing primary product prices or export earnings of less developed countries has to be examined for each country separately on the basis of an analysis of its economic structure" (p. 372). Other studies working with much more detailed country models include work by Ricardo Lira on copper in Chile, by Pedro Palma on iron ore and petroleum in Venezuela, by Adams and Roldan on Brazil and coffee, and by Paul Acquah on Ghana and cocoa. Only the last of these studies is explicitly focused on the impact of the world commodity market on growth of the producing economy. It showed that fluctuations in the world cocoa market generated internal fluctuations in the Ghanaian economy and that if these fluctuations could be smoothed, the growth potential would be increased.

An essential aspect in this type of work is to provide more than a broad aggregative link between the world commodity market and the country's economy. In the context of the country macro model, the specific commodity producing sector must be broken out and the linkages to the other sectors of the economy must be represented in the model structure. While the theory of sectoral behavior and linkage is not complex, the data requirements may be hard to meet. Needed are detailed statistics, preferably in time-series form at the sector level. Even when these are available, they require reconciliation with aggregate statistics. Often the timing will be different (crop years rather than calendar years), the units will vary (bags of coffee as compared to deflated value data), and trade statistics may not correspond (customs data as compared to volume export data, for example). Frequent-

*University of Pennsylvania and Wharton EFA, Inc

ly, short time-series, cross-section data, or engineering estimates must be used. In all cases, the commercial, technical, and government regulation aspects must be known and taken account of in the model.

Another important consideration is the recognition that the relationships between the country model and the world commodity market are two-way linkages. If the producing sector is relatively important in the world commodity economy, there will be feedbacks from the country to the world commodity market—frost in Brazil affects the Brazilian harvest and influences the world price of coffee. Indeed, such an event may be beneficial to Brazil—the 1975–76 frost sharply increased Brazilian coffee export earnings in the following years. The two-way interaction also is likely to have important dynamic aspects—a price rise today will influence supply and price many years later. Some phenomena may of course be studied on a one-way basis, assuming that the world market impacts on the domestic sector, but others require the capability for two-way interaction between the producing country model and the world commodity market.

The potential for domestic policy response is another complicating element. There are probably few cases where a producing country can greatly influence the world commodity market with unilateral policy actions—the case of Brazil and coffee is said to be such a case. But in the domestic economy policy intervention may be the rule rather than the exception, so that domestic policy such as export taxes, neutralization of foreign exchange earnings and trade restrictions may be used to offset commodity market fluctuations. Therefore, observed patterns of impact on the domestic economy may be quite different than would be apparent from simple simulations which disregard government intervention.

II. Linkages of the Coffee Market to the Brazilian Economy

The conceptualization of the linkages between a commodity producing sector and the macro economy is relatively straightforward. We will limit ourselves here to an enumeration of the main impacts.

The principal impacts of the coffee sector on the macro economy of Brazil can be summarized as follows: (We include the direct links as well as the feedbacks.)

The production-income link. Variations in coffee output directly affect the gross domestic product of the economy, disposable income, and the corresponding flows of wage and nonwage income.

Production-employment and input linkages. Additional coffee production raises demand for labor and inputs from other productive sectors of the economy.

Export value balance-of-payments effect. The value of exports (resulting from changes in volume exported or international price) significantly affects the Brazilian balance of payments and, consequently, the ability to import.

Tax revenue. Taxes are imposed on coffee exports and their impact on the Coffee Fund revenues is important. The latter disburses part of its revenues to the government where they may affect public spending.

Monetary effects. Monetary effects occur, through the impact of coffee earnings on foreign exchange receipts and the net balance of the Coffee Fund operations.

Consumption effects. Consumption demands originate in the flow of private and public revenue.

Investment effects. Coffee activity affects the level of aggregate investment by various channels. Government revenues flow into government investment expenditure. Foreign exchange earnings allow increase in imports of capital goods and impact on investment. The growth of Gross Domestic Product (*GDP*) calls for higher levels of investment spending. A critical issue to be considered below is the effect of additional investment on the supply potential of the developing economy.

Wage and price effects. Potentially wages play a role since the commodity producing sector could be a leading sector in the process of wage determination. High wages paid in some primary producing sectors have contributed an upward bias to wages in

some developing countries. In Brazil, coffee wages are largely determined by minimum wage regulations. Prices are also affected by monetary forces, an important feedback link in the Brazil model.

An important issue in this research is the relative role of forces on the demand side and on the supply side in determining the development path. Does the impetus come from demand, automatically drawing forth the supply-side response? Or, alternatively, are the supply-side constraints operational so that they must be modeled explicitly? The theory of economic development would suggest the latter. However, whether as a consequence of modeling practice or the reality of underutilization in developing country industry, the emphasis in econometric modeling has been on the demand forces. These involve the backward linkages from the commodity producing industry, the requirements for products of tertiary industry, and the demand for consumer and investment goods generated by the income flows. On the supply side the main effect is through public investment, presumably in infrastructure, which increases primary sector output, and to a much lesser extent through private investment. The supply-side forces which should have dynamic impact on the growth potentials of the economy are poorly represented, relative to demand side phenomena, a fact which is apparent both from model structure and from multiplier simulations. There is no conclusive answer as yet whether the demand-side influences force an appropriate supply-side adjustment or whether it will be necessary to recognize the supply-side constraints more explicitly than at present.

III. Simulation Results

A number of experiments have been carried out with a commodity-country model system to illustrate different aspects of the interaction between the commodity producing sector, the macro economy, and the world commodity market. The commodity model is a detailed representation of the world coffee market with endogenous explanation for production, consumption, inventories, and price. The country model linked to the coffee model is an adaptation of a fairly detailed macro-forecasting model of Brazil. This is part of a broader project which will see construction of a linked system of models— Brazil, Ivory Coast, and Central America and the world coffee market and, similarly, Chile and Zambia and the world copper market.

Two computations involve an exogenous increase in Brazilian coffee yield. In Case 1 we focus on domestic effects by assuming that exports are set in the world market and that the additional production is absorbed in Brazil. Most of it ends up in private inventories. Case 2 adds the foreign trade effects by assuming that additional coffee output is exported to the world market (though we assume no change in the world coffee price).

The results of this computation are summarized in Table 1. It has been assumed that over the period 1976–87 yields would be 10 percent higher than in the base forecast. Effects are then computed as the percentage impact relative to the base solution value. The extent of the linkage between the coffee sector and the Brazilian economy is summarized by the impact on *GDP*, amounting to 0.4 percent when coffee production increases by 10 percent. Coffee production accounts for roughly 1.7 percent of *GDP* during the sample period. Thus the direct impact of coffee production of *GDP* is less than 0.2 percent, so that the indirect repercussions account for a substantial portion of the total effect. The main channels of impact are the needs for transport and commercial services, the higher government revenues and expenditures, and the impact of growth of demand on investment.

If we assume that the larger volume of production goes into exports rather than into domestic inventories, and that there is no change in world price, then Brazilian coffee exports would increase by 17 percent during the simulation period, an increase in export revenues of 2.2 percent on average. As the foreign trade variables are allowed to operate in this solution, the main channel through which they affect the domestic economy is through the increased capacity to import and the increase in government revenues accruing through the coffee export tax. The impact is

TABLE 1—AVERAGE PERCENTAGE DIFFERENCES BETWEEN DISTURBED AND BASE SOLUTIONS; 10 PERCENT HIGHER COFFEE PRODUCTION OVER THE BASE SOLUTION; DYNAMIC SIMULATION, 1976–87

		Domestic Effect Only[a]	Domestic and Foreign Effect[b]
GDP	Gross Domestic Product	0.44	0.81
PGDP	Deflator Gross Domestic Product	−0.05	0.77
CPR	Real Private Consumption	0.50	1.27
IPR	Real Gross Private Investment	0.45	0.74
CGR	Real Government Consumption	0.34	0.55
IGR	Real Government Investment	0.56	0.74
X2R	Industrial Output	0.36	0.72
X3R	Services Output	0.40	0.80
TDC	Direct Taxes	0.59	1.54
TINDC	Indirect Taxes	0.36	1.54
LCF	Coffee Sector Employment	3.69	4.68
VACFC	Coffee Producers Price	−7.47	4.72
MKCIFR	Capital Goods Imports, real terms	0.98	2.01
ECF	Coffee Export, real terms	0.0	17.43
S/STOCKCF	Ratio of Coffee Sales Over Stocks	−15.85	9.09

[a]Assuming added coffee production is not exported.
[b]Assuming added coffee production is exported, but has no impact on the world coffee price.

considerable. The *GDP* is higher on average by 0.81 percent compared to the 0.44 percent effect without the increase in exports. A particularly notable repercussion is the increase in capital goods imports of 2.01 percent. The inclusion of foreign balance effects also has a significant effect on inflation by way of the money supply. This price impact materially changes the economic situation of the Brazilian economy.

The other model simulations deal with the impact of external coffee market forces on the Brazilian economy. In Cases 3–5 we have assumed that a blight affecting African coffee production resulted in coffee prices on average 26 percent higher than in the base solution. The volume of Brazilian coffee exports increases by 8 percent on average so that export revenues from coffee increase 34.6 percent. Total export revenues are higher by 4 percent on average over the base solution figures. In Case 3 we assume no change in the coffee export tax. The higher international coffee prices and the increase in export sales drive producer prices up by 10.7 percent, moderately reduce domestic coffee consumption and eventually increase production. The net positive impact derived by Brazil from this set of events is a 0.31 percent increase in *GDP* relative to the base solution. The price level is 1.71 percent higher.

In order to test the impact of export tax policy, we have carried out the same simulation assuming that export taxes are raised sharply from a negligible level in Case 3 to 15 percent of the coffee export price in Case 4. Consequently, domestic coffee prices increase less rapidly, consumption decreases less, and the expansion of production is moderated. The result is a smaller impact on *GDP*, only 0.23 percent as compared to the base run, and a higher inflation rate, 2.49 percent higher than the base run.

Finally, we have recognized that the increase in coffee exports will result in higher inflation. To offset this potential inflationary impact, we also have examined a case in which the monetary expansion would be neutralized, using monetary policy to offset the increase in reserves. This results in a lower rate of inflation (only 1.00 percent higher prices on average than the base run) and a higher real *GDP* as compared to the simulation without policy intervention or the one which only imposes an export tax policy (.43 percent higher than the base run).

IV. Conclusions

The computations described in this paper indicate that fluctuations in the coffee market have magnified impacts on the macro variables of the Brazilian economy. Increases in coffee output and/or in the value of coffee exports translate into a higher real product with effects on income, government revenues, secondary and tertiary sector activity, investment, foreign exchange earnings, imports, etc. Prices are also affected.

In order to establish the nature and magnitude of these effects, we have modeled the coffee sector and traced through the linkages from this sector to other parts of the Brazilian economy. These computations suggest that the principal linkages operate through: 1) the effect of coffee production on requirements for secondary and tertiary sector output; 2) the impact of earnings from coffee production, both wage income and nonwage receipts, on consumer demand; 3) the impact of coffee export tax payments on government revenues and subsequently on government spending for consumption and investment; 4) the effect of increased foreign exchange earnings (in cases where value of coffee exports increases) on imports and subsequently on capital formation; 5) the effect of increased foreign exchange earnings on money supply and prices.

The predominant effects in the present version of the model system are through the demand linkages, though there is some supply impact by way of capital formation. The predominance of the demand effect is not altogether surprising. The Brazilian macro model does not contain capacity constraints on industrial and tertiary sector output, reflecting the significant industrial underutilization which exists in Brazil as in many other developing economies.

An important task which remains is to elaborate the linkages on the supply side and to measure their full dynamic impact on economic development.

REFERENCES

P. Acquah, "A Macroeconometric Analysis of Export Instability in Economic Growth," unpublished doctoral dissertation, Univ. Pennsylvania 1972.

F. G. Adams and R. Roldan, "Economic Studies of the Impact of Primary Markets on Economic Development in Latin America," paper prepared for the Nat. Bur. Econ. Res. Conference on Commodity Markets, Models and Policies in Latin America, 1977.

R. Lira, "The Impact of an Export Commodity in a Developing Economy: The Case of the Chilean Copper: 1956–1968," unpublished doctoral dissertation, Univ. Pennsylvania 1974.

P. Palma, "A Macroeconometric Model of Venezuela with Oil Price Impact Applications," unpublished doctoral dissertation, Univ. Pennsylvania 1977.

C. Rangarajan and V. Sundararajan, "Impact of Export Fluctuations on Income: A Cross Country Analysis," *Rev. Econ. Statist.*, Aug. 1976, *58*, 372.

Robust Stabilization Policies for International Commodity Agreements

By BRUCE GARDNER*

This paper reviews the relationship between inventory management and price stabilization and then considers optimal market control schemes under imperfect knowledge. Emphasis is placed on strategies that avoid costly errors even if formulated using incorrect information.

I. Storage Rules and Price Stabilization

Consider a grain market taken to have the following characteristics: the commodity is produced in an annual cycle, is storable, has relatively inelastic demand and supply functions, and a substantial random element in production. These characteristics suggest large year-to-year price variability in the absence of stockpiling, and the desirability of year-to-year storage. The sequence of events starts at the beginning of each marketing year when the crop is harvested. The harvest plus carryin (beginning) stocks from previous crops is the current supply. The current supply must be allocated between consumption during the current year and carryout (ending) stocks for use in the future. Ending stocks plus next year's randomly distributed harvest are next year's supply.

Because the choice of the quantity held in ending stocks determines quantity allocated to current consumption, the storage decision determines market price as the price at which the quantity remaining for current consumption clears the market. Given the value of the state variable, supply, all the endogenous variables—ending stocks, consumption, and price—are determined simultaneously. Thus, the control variable could be specified as consumption or as price instead of ending stocks. Diagrammatically, ending stocks at each price represent a reservation demand for grain stocks which may be added to current consumption demand to obtain a total demand (see the author, p. 78). (In a private market context, the form of the reservation demand for ending stocks provides a good structural basis for the non-linear price equation in models such as F. Gerard Adams and Jere Behrman.)

Despite the obvious elements of duality between stockpiling and price stabilization, the nature of the relationship between optimal storage and price stabilization has not been made clear in the literature. Indeed, there exist two largely independent areas of research, one on optimal inventory control (stemming from Robert Gustafson), the other on optimal price stabilization (with its most recent impetus from Benton Massell).

Optimal inventory control involves the discovery of a storage rule, which specifies ending stocks as a function of current supply and other variables. The optimal storage rule maximizes the expected value of the appropriate objective function, taken in the present work as the sum of current and expected future consumer's and producer's surpluses minus expected net external losses due to future shortfalls in grain availability. In the absence of externalities, optimal storage would be that undertaken by a risk-neutral private competitive storage industry.

The literature on the welfare costs of price instability provides analyses of expected consumer's and producer's surplus for a given standard deviation of price or quantity. Massell demonstrates a welfare gain to reducing price variability if stabilization is costless. Extensions of his approach have not to date been successful in deriving optimal storage rules in the presence of costly storage. The only practical tool for welfare analysis applied to date is simulation. However, most of the available simulation studies do not simulate stockpiling regimes derived by optimizing methods. (For a notable exception, see Hovav Talpaz and C. R. Taylor.)

The stochastic element of production

*Texas A&M University.

generates a random series of observed supplies from which the optimal storage rule determines stocks, consumption, and price. The probability distribution of production determines the probability distribution of these other variables. The probability distribution of current price is derived from the probability distribution of consumption via the demand function. The variance of this price distribution is optimal—it is the variance associated with maximum expected value of the objective function. Even if the probability distribution of production has some standard form such as normal or triangular, the probability distributions of supply, consumption, and price will not in general have any simple form.

II. Optimal Market Management: Robust Policies

In analyzing storage and price stabilization in the context of an international commodity agreement, it is not necessary that every country participate or subordinate its domestic commodity policy entirely to the international agreement, or that all commodities related to the main commodity of interest be covered (for example, wheat could come under an agreement, while feed grains do not). But it is necessary that the agreement determine "offshore" prices in world trade, and that domestic policies provide a relatively stable environment in which the international agreement may function.

The optimal storage rule and corresponding price stabilization policy depends upon the estimated elasticity of demand, the elasticity of production in response to expected price, the costs of storage (including interest), the probability distribution of yields, and an assumed schedule of external costs of consumption shortfalls. Two simple but important general features of optimal storage rules are: 1) there is no such thing as an optimal "reserve stock level" apart from the situation in a particular year; and 2) the optimal stabilization policy should not be based on price bands (a reservation demand for stocks which becomes perfectly elastic at a floor price and again at a ceiling price).

What is the likely magnitude of welfare loss from following a suboptimal storage policy? This question is not as easy to answer as might at first be supposed. While simulation of an optimal stockpiling regime is straightforward, simulation of suboptimal policies is complicated by the fact that these policies often leave scope for private speculative storage. I have discussed the proper modeling and simulation of such policy regimes elsewhere Generally, private stockpiling tends to fill in gaps left by, for example, an International Commodity Agreement (*ICA*) buffer stock with a wide price band, and by doing so reduces the welfare losses that would otherwise result from this policy.

Table 1 shows results from a series of experiments with suboptimal but easily operational stockpiling strategies, including buffer stocks of several different maximum sizes and different acquisition and release rules. For each policy regime it is assumed that private speculative stockpilers know the policy and engage in storage to maximize expected profits given the policy. The first row of Table 1

TABLE 1 – EXPECTED WELFARE UNDER ALTERNATIVE STORAGE STRATEGIES

	Optimal Storage Rule	Ignore Externalities[a]	Small Buffer Stock[b]	Large Buffer Stock[b]
Base Case	100.0	99.5	99 9	99.5
No Externalities Exist	100.4	100 7	100.4	99.9
Externalities Greater than Estimated	98.0	95.8	98.1	97.7

[a]Equivalent to privately optimal storage by competitive speculative stockpilers

[b]The "small" buffer stock has an upper limit of 8 percent of mean production, and the "large" buffer stock has an upper limit of 15 percent of mean production. Both have an acquisition price of 0.9 times mean price and a release price 1.6 times mean price.

compares the mean value of the objective function (consumer's plus producer's surplus minus externalities) under optimal storage to the results of three suboptimal storage strategies for the world wheat market. The scaling is such that 100 equals approximately the mean expenditures on wheat, roughly $40 billion worldwide, so that a one point reduction in welfare amounts to about $.4 billion.

Optimal storage policy is derived from estimated parameter values that are consistent with existing empirical studies of the wheat market, but which are quite likely to be incorrect. Our knowledge is especially weak concerning external costs associated with consumption shortfalls. The nature of such externalities is discussed, but no estimates given, in John Stein and Rodney Smith. The schedule of externalities assumed in deriving the "base" storage rule for Table 1 implies that the marginal social value of wheat increases roughly 20 percent for each 1 percent reduction below mean consumption. In about 3 years out of 100, a production shortfall would be experienced which would generate worldwide external costs of 25 percent or more of the mean value of wheat expenditures (i.e., $10 billion), if no reserve stocks were available.

The second row of Table 1 shows indexes of expected welfare when external costs are neglible, but storage policy is carried on as if they were not. The "optimal" strategy from the first row is now suboptimal. The new optimal strategy is to ignore externalities (as the private trade would do). The third row of Table 1 shows indexes of expected welfare when external costs are larger than assumed deriving the original storage rule, but storage policy is carried out under the original rule. In row 3, a production shortfall which would generate external costs of $10 billion or more if no reserve stocks were available would occur in 1 out of 10 years. None of the row 3 strategies is exactly optimal, but of those shown, the small buffer stock is best.

A lack of confidence in estimated externalities, as well as the potential for error in estimates of wheat demand, supply, and storage cost functions, encourages a search for storage strategies which are *robust*, that is, which are not too suboptimal over the range of likely error in our knowledge. A storage policy which is suboptimal if the externalities and market parameters are exactly as assumed might nonetheless be superior to an apparently optimal policy which is sensitive to error in estimated externalities or market parameters. A crude test for robustness in the Table 1 results is to calculate mean expected welfare for each stockpiling strategy for the three situations considered. More generally, one should use the (subjective) probability distribution of the size of externalities as well as the probability distribution of the market parameters which determine the optimal storage policy to calculate an expected welfare index for each proposed storage policy, that is, a much larger version of Table 1.

The overall results of the experiments to date are that while socially optimal stockpiling rules tend to be fairly robust, some buffer stock regimes are sufficiently robust to be as good as or even preferable to an optimal rule strategy when a plausible range of uncertainty about the parameters used in deriving the optimal policy is taken into account. However, it is important to exercise care in the choice of buffer stock regime. A striking result of the experiments is that the upper limit on the buffer stock tends to be as important as the price band used. The economic reason is that too large a stock can easily run up storage costs which at the margin provide no corresponding benefits, while a small maximum size leaves the system open to large external costs from shortfalls. With respect to price bands, generally preferable results were found with price bands centered substantially above mean expected price (although this makes it even more important not to get the maximum buffer stock size too high). This result may seem surprising because an optimal storage rule typically conducts most of its acquisition and release of stocks at prices below mean price, with very small if any stocks held when price is above mean price. The reason why the best buffer stock price band is higher is that private speculative stocks tend to operate in the lower range of prices. The main function of the buffer stock is to hold grain off the market for use when the extreme shortfalls occur for which private stocks are likely to be

insufficient. A high release price holds the reserve stocks until the externalities become large. A relatively high acquisition price helps insure the availability of these stocks.

The limitations of these experiments should be recognized. They do not bear on the issues of whether externalities in fact exist which warrant intervention, or whether an *ICA* is actually feasible or desirable. The issue here is the more narrowly technical question, given that an *ICA* or other stabilization scheme using stockpiling has been decided upon, how it might best be formulated in the presence of externalities and imperfect knowledge.

REFERENCES

F. Gerard Adams and Jere R. Behrman, *Econometric Models of World Agricultural Commodity Markets*, Cambridge, Mass. 1976.

B. L. Gardner, "Optimal Stockpiling Models and the Effects of Public on Private Storage," *Proc. Amer. Statist. Assn.*, Aug. 1977, 77–83.

R. L. Gustafson, "Carryover Levels for Grains," tech. bull. no. 1178, U.S. Dept. of Agriculture, Oct. 1958.

B. F. Massell, "Price Stabilization and Welfare," *Quart. J. Econ.*, May 1969, *83*, 285–97.

J. P. Stein and R. T. Smith, "The Economics of United States Grain Stockpiling," Rand Corp., R-1861, Mar. 1977.

H. Talpaz and C. R. Taylor, "Optimal Wheat Reserves in the United States," *Amer. J Agr. Econ.*, forthcoming.

The New Jobs Tax Credit: An Evaluation of the 1977–78 Wage Subsidy Program

By JEFFREY M. PERLOFF AND MICHAEL L. WACHTER*

The New Jobs Tax Credit was one of the four programs in the 1977 economic stimulus package. This program, although viewed primarily as a countercyclical measure, may also alter the equilibrium unemployment rate, U_N.[1] This paper presents our preliminary analysis of the Department of Labor survey, conducted by the Bureau of the Census, in which firms described their responses to this employment tax credit (*ETC*). To date, our results indicate the potential for a large employment effect. Ordinary least squares estimates suggest that firms which knew about the program increased employment 3 percent faster than other firms. A second analysis which uses multinomial logit techniques indicates that the *ETC* shifted the entire distribution of employment growth to the right: slowly growing firms increased employment to capture the credit. Since the firms which knew about the program, however, were not randomly drawn, our results may overstate the program's employment effect.

Due to the nature of the survey data, we can only focus on direct employment effects.[2] It is useful, however, to at least mention the other potential effects of the program. First, unlike Comprehensive Employment Training Act (*CETA*) programs which increase public employment, the *ETC* should increase employment in the private sector. Within the private sector, the rules of the current *ETC* program provide an additional stimulus to the growing industries and, to a lesser extent, to small establishments. Second, the long-run structural effects of this two-year program are probably small. A permanent *ETC*, however, may be able to lower U_N of disadvantaged workers.

I. The New Jobs Tax Credit

If the economy is in equilibrium, the micro impact of a wage subsidy is straightforward. The exact effect, however, depends upon the specific type of wage subsidy. For a more detailed description of the New Jobs Tax Credit than given below, see Orley Ashenfelter and the Congressional Budget Office.

The *ETC* of 1977 has four key provisions: First, the subsidy is paid to employers rather than employees. As a result, it shifts the derived demand curve rather than the supply curve of labor. By lowering the wage rate, the *ETC* induces a shift away from capital towards labor within the firm and from capital to labor intensive firms.

Second, the *ETC* gives a greater subsidy for unskilled and part-time labor than for skilled or full-time workers. The tax credit is

*Assistant professor and professor of economics, respectively, University of Pennsylvania. We wish to thank D Nichols and H Pitcher in the Labor Department, P Ohs and J McDonald in the Bureau of the Census, and M Asher at the University of Pennsylvania for extensive help. Hugh Pitcher, who was involved in all stages of the project, did much of the initial programming of the Cross-Section Processor package which was used to estimate multinomial logits Our ongoing research in this general area has been supported by the Department of Labor, the General Electric Foundation, and the National Institute of Child Health and Human Development

[1]See Wachter (1976) for a derivation of U_N. The normalized unemployment rate is an approximation to the nonaccelerating-inflation rate of unemployment and captures shifts in the demographic composition of the labor force and change in relative unemployment rates.

[2]The stress in evaluation studies is usually placed on calculating the dollar cost per new job created as a *direct* result of the program. Other factors should also be investigated. For example, see Perloff and Wachter (1978).

limited to the minimum of either 50 percent of the excess total wages over 105 percent of the previous year's total wages or 50 percent of the excess of wages covered by federal unemployment insurance (*FUTA* wages)[3] in the 1977 or 1978 tax year over 102 percent of *FUTA* wages from the previous year. Although the 105 percent of the increase in total wages condition prevents firms from firing all their full-time workers and hiring twice as many half-time workers, it does not prevent them from hiring part-time incremental workers. Further the credit may be no larger than the smaller of 25 percent of *FUTA* wages or $100,000. Thus the credit (C) is

$$\begin{aligned} C &= \min\,[0.5(4200)(E - 1.02E_0); \\ &\quad 0.5\,(WE - 1.05W_0E_0);\ 100{,}000; \\ &\qquad 0.25(4200E)],\ W,\ W_0 \geq 4200, \\ &= \min\,[0.5(WE - 1.05W_0E_0); \\ &\qquad 100{,}000;\ 0.25WE],\ W,\ W_0 < 4200, \end{aligned}$$

where E is current employment, E_0 is employment in the previous year, W is the current wage, and W_0 is the previous year's wage.

Third, the *ETC*, as indicated by the formula above, is clearly a marginal credit affecting incremental hirings rather than a wage subsidy covering total employment. A firm which, given the credit, chooses to grow by an amount within the subsized range has its marginal wage affected. The proportion by which the wage is reduced (θ) is

$$\begin{aligned} \theta &= 0 && E/E_0 < 1.02 \\ \theta &= 2100/W && 1.02 < E/E_0 < E^*/E_0 \\ \theta &= 0 && E/E_0 > E^*/E_0 \end{aligned}$$

where E^*/E_0 defines the employment ratio above which the marginal subsidy vanishes. Ashenfelter shows that $E^*/E_0 < 2.04$.

Under the current program, firms in a rapidly growing industry essentially receive a lump sum subsidy which leaves the marginal wage unaltered. Such a subsidy is similar to a negative franchise tax in that it lowers the long-run average cost curve of each firm. If the industry is made up of identical firms in sequential, short-run competitive equilibrium, industry employment will increase as new firms enter (but the output of any one firm decreases). Thus, the current program might affect employment in certain growth industries even if it does not affect the marginal wage.

Fourth, the program has important dynamic properties since it is only a two-year program and a new worker only generates wage savings for one year. The one-year coverage may generate a sawtooth employment pattern. A long-run New Jobs Tax Credit program might induce a firm with constant labor requirements to increase employment the first year and inventory its extra output, then decrease employment in the second year in order to qualify for the credit in the third year, and so forth. Short-term employees under such a program may receive little on-the-job training and hence little long-run benefit.

The macro-economic impact of the current complex *ETC* is extremely difficult to model. Indeed there have been few attempts at modelling even simple wage subsidies in a macro-economic, disequilibrium model (for example see Gary Fethke and Samuel Williamson). Within a Barro-Grossman type disequilibrium model, in which firms are off their notional demand curves but on their effective demand curves and there is disequilibrium in the labor market associated with an incorrect *nominal* wage, a properly financed subsidy can help adjust the nominal wage to correct the disequilibrium. This scenario assumes, however, that the government is "foolish" enough to allow the disequilibrium to occur in the first place, but "smart" enough to find the new market-clearing wage. Alternatively, if the money supply is adjusted to offset the subsidy, then the subsidy will not have a beneficial effect. In any case, since the relative factor input costs are affected, there will be an incentive for firms to operate at a different (and probably inefficient) combination of inputs. It can be shown that an adjustment of the money supply will always dominate the employment tax credit as a means of correcting this type of disequili-

[3]As of January 1, 1978, the *FUTA* maximum went up to $6,000, but the Jobs Tax Credit only subsidizes the first $4,200 (the 1977 minimum) per employee.

brium, both in terms of administrative ease and efficiency in production.

The *ETC* program may also have positive long-run employment effects, reducing the equilibrium unemployment rate U_N by lowering the real wage paid by employers. The *ETC*, rather than merely offsetting the investment tax credit (*ITC*), may work with the *ITC* to increase employment by reducing U_N. Both programs can provide an offset to minimum wage and government transfer (for example, Aid to Families with Dependent Children) effects.

II. Descriptive Statistics

Our empirical study of the New Jobs Tax Credit is based on a two-page Department of Labor financed, Bureau of the Census questionnaire mailed to 4,266 for-profit firms in February 1978. This study used the 1,473 firms which answered all the key questions. To maintain confidentiality, the survey results have not been merged with other tax data, so the only information available is this two-page questionnaire.

The descriptive statistics (see U.S. Department of Commerce) indicate that relatively few firms knew about and responded to the *ETC*. Of the weighted sample of firms who responded in full (weights compensate for the oversampling of large firms) 34.4 percent know about the *ETC*. Only 27.3 percent of firms with 0–9 employees knew about the *ETC* compared with 89.1 percent of firms with over 500 employees. Thus, large firms were almost three times as likely to know about the credit as small firms.

Of the firms who knew about the credit, 19.5 percent believed they qualified for the credit while approximately one-quarter did not know if they qualified. Almost three out of ten of the firms (especially small firms) who knew of the *ETC* said they did not qualify because of insufficient growth of *FUTA* wages. Insufficient growth of total wages disqualified almost one-fourth of the firms (especially small firms). Approximately 3 percent of the firms (especially large firms) did not qualify because of insufficient tax liability.

Of the firms who knew about the credit, only 6.1 percent said they made a conscious effort to increase employment. Thirty-six percent said their employment growth was so high that they automatically qualified for the credit, 7.2 percent learned of the credit too late, 5.0 percent said that the credit was too "troublesome to pursue," and the rest gave other reasons or did not answer the question.

To conclude, relatively few firms knew about the program and even fewer firms claimed to have made a conscious effort to respond to the program. The National Federation of Independent Business survey of 2.2 million firms produced similar descriptive statistics (see Robert Eisner).

III. Empirical Results

Since only certain firms knew about the *ETC*, one is tempted to treat the survey data as though it were produced by a large experiment with only certain firms exposed to the *ETC* stimuli. While we gave in to this temptation, the following results should be viewed with caution: the firms which knew about the *ETC* were not randomly chosen. It is possible, for example, that a variable such as entrepreneurial skill is responsible for both knowing about the program and rapid growth. Similarly, firms with rapid employment growth had an incentive to learn the specifics of the program.

Due to data limitations, the effects of the *ETC* cannot be analyzed within a structural model of a firm's decision-making process. For example, we cannot be certain as to which of the New Jobs Tax Credit's restrictions were binding for particular types of firms. Given these limitations, we employed a quasi-reduced-form specification.

Table 1 presents ordinary least squares regression results which utilize the questionnaire data. The dependent variable in all equations is the percent change in employment between 1976 and 1977.[4] Virtually all available measures are used as independent

[4]For reasons of confidentiality, data were supplied to us in terms of dummy variables or percents. No levels data were provided. The use of size dummies (rather than a continuous variable) introduces an errors-in-variables problem.

TABLE 1—PERCENT INCREASE IN EMPLOYMENT: ORDINARY LEAST SQUARES
(Standard errors in parentheses)

Independent Variable	Sample: All Firms I	II	III	IV	Firms Which Knew About The ETC V	VI
Constant	-25.2	-7.88	9.56	37.1	5.37	27.43
	(9.28)	(11.0)	(9.96)	(11.8)	(10.7)	(13.34)
Firms:						
1-9	2.50	4.40	1.54	2.46	0.283	1.23
	(1.67)	(1.99)	(1.56)	(1.85)	(1.85)	(2.33)
10-49	3.81	4.26	1.69	2.51	2.35	2.88
	(1.51)	(1.79)	(1.46)	(1.74)	(1.89)	(1.99)
50-249	2.24	2.89	1.04	1.89	1.25	2.20
	(1.51)	(1.80)	(1.49)	(1.78)	(1.53)	(1.93)
250-499	0.772	0.184	0.381	0.983	0.391	-0.788
	(1.98)	(1.89)	(1.59)	(1.40)	(1.57)	(1.97)
Industry:						
Manufacturing	1.17	1.62	0.639	0.911	1.46	2.36
	(1.37)	(1.57)	(1.34)	(1.59)	(1.47)	(1.85)
Trade	1.08	-1.28	0.865	1.48	1.55	0.548
	(1.20)	(1.45)	(1.21)	(1.44)	(1.49)	(1.87)
Service	0.780	0.271	0.477	0.186	3.95	3.25
	(1.38)	(1.62)	(1.37)	(1.62)	(1.64)	(2.05)
Tax Form:						
1040	1.52	1.08	1.22	0.633	2.56	1.84
	(1.38)	(1.65)	(1.39)	(1.65)	(2.01)	(2.53)
1065	-0.951	2.21	1.08	-2.31	1.26	0.496
	(2.06)	(2.45)	(2.08)	(2.47)	(2.57)	(1.21)
Region						
NE	0.887	2.14	0.515	2.47	1.76	-1.85
	(1.29)	(1.53)	(1.30)	(1.54)	([illegible])	(1.91)
NC	0.228	1.84	0.369	-1.97	2.30	3.25
	(1.27)	(1.52)	(1.28)	(1.82)	(1.52)	(1.91)
South	0.970	-0.973	0.495	1.44	-0.952	1.03
	(1.24)	(1.48)	(1.25)	(1.49)	(1.52)	(1.91)
% Change in Sales	0.495		0.492		0.530	-
	(0.20)	---	(0.0201)		(0.0247)	-
Know of ETC	3.07	3.53				
	(0.934)	(1.11)	-	--	--	
Conscious Effort to Increase Employment	-	-	10.26	14.7	9.13	11.8
	-		(3.21)	(2.62)	(1.98)	(2.47)
R^2	0.31	0.014	0.31	0.029	0.40	0.058
SSR	388,114	551,534	379,317	517,365	143,117	[illegible]
SE	16.37	19.44	16.39	19.38	14.31	17.99
DF	1458	1459	1430	1431	795	796

variables in the equations. For our purposes, the key independent variables in these equations are the knowledge and conscious effort dummies. The first two regressions use a knowledge dummy which is equal to one if the firms knew about the *ETC* and zero otherwise. The coefficient on the knowledge variable is statistically significant at the 0.05 level in both equations. The main implication of these equations is that firms which knew about the credit increased employment by over 3 percent more than similar firms that were ignorant of the program. Due to the design of the program, however, there is an incentive for firms to hire part-time workers at the margin (since only *FUTA* wages count for the credit). In these regressions, we cannot distinguish between part-time and full-time workers.

In the remaining equations, a dummy variable which has the value of unity if a firm made a "conscious effort" to increase employment and zero otherwise is used instead of the "knowledge" of the *ETC* variable. This variable is statistically significant in all the equations in which it appears. Firms which made a conscious effort increased employment by over 10 percent more than all other firms and by over 9 percent more than firms which knew about the *ETC* but made no special effort to qualify.

Other independent variables are included in these regression equations to explain the percent change in employment. Regional and industry dummies, included to reflect changes in demand, are not statistically significant. Some firms' size dummies are statistically significant at the 0.05 level.[5]

The percent change in sales was included in three of the six equations in Table 1. An increase in sales could be expected to cause a somewhat smaller increase in employment since some of the increase in sales was due to inflation between 1976 and 1977. This variable has a coefficient of roughly one-half in all the equations and is statistically significant. Equations which include sales growth have R^2s of 0.3 or 0.4, whereas other equations have R^2s below 0.06. Although this variable is endogenous, a lack of appropriate instruments precluded the use of instrumental variables techniques.

The implication of these linear regressions is that knowledge of the program or a conscious effort to respond to the *ETC* causes a firm to have a greater increase in employment. Again, some of these findings may be the incidental result of the nonrandom division of firms into knowledgeable and unknowledgeable firms. While the equations which only include knowledgeable firms cannot be faulted on these grounds, the decision to make a conscious effort to increase employment depends on employment growth (due to the restrictions of the program) and hence the conscious effort variable is endogenous.

[5]Since many large firms may collect the $100 thousand maximum credit and not have their marginal wage altered, we tried knowledge and firm size interactive dummies. The hypothesis that knowledge affects employment growth in all size firms equally cannot be rejected at the 0.05 level.

The complicated rules of the New Jobs Tax Credit renders suspect the linear reduced-form specification used above. Rather than affecting all knowledgeable firms equally, the *ETC* should only influence firms in a fairly narrow growth range. For example, a firm which planned to contract employment substantially would not be eligible for the program because of inadequate employment growth. Similarly, a firm with very rapid growth would qualify for the *ETC* automatically and not have its marginal wage altered. The regression results, however, do not allow for different types of effects across firms.

The multinomial logit approach does allow for different size *ETC* effects across the employment growth distribution of firms. In addition, it permits non-linear firm size interactions. Finally, it can fit the employment growth distribution which cannot be fitted by a parametric function. For example, the unconditional frequency distribution has a large spike at zero. Roughly one-third of all firms and close to half of the small firms reported no change in employment from 1976 to 1977.

In the following multinomial logit framework, the hypothesis to be tested is that the distribution of employment growth of firms is shifted to the right by the *ETC*. This analysis suggests that, on balance, firms which were not growing or were growing slowly increased their employment growth in order to receive the credit, while firms with rapidly declining employment were relatively unaffected.

To test this hypothesis, we conducted the following experiment. The distribution of rates of growth of employment was broken up into intervals: $S_1 = (-\infty, -1\%]$, $S_2 = (-1\%, 2\%]$, $S_3 = (2\%, 30\%]$, $S_4 = (30\%, 45\%]$, and $S_5 = (45\%, \infty)$. The multinomial logit approach allows us to estimate the probability of a firm being in one of these mutually exclusive intervals.[6]

Table 2 presents our multinomial logit results. The *log* odds of being in one of the above mutually exclusive intervals ($ln\ P_j/P_1$, $j = 2, 3, 4, 5$, where P_j is the probability of firms being in interval S_j) are estimated using S_1 as the base category whose coefficients have been normalized to zero. The first multinomial logit is restricted to firms with between zero and 49 employees, and the second includes firms with 50 or more employees.

Not all coefficients on the knowledge variable are asymptotically significant. Since the probabilities must sum to one, if even one coefficient differs from zero, it can affect all the probabilities; weight shifted to one part of the distribution must come from elsewhere. For both large and small firms, likelihood

TABLE 2—MULTINOMINAL LOGIT DISTRIBUTION OF EMPLOYMENT GROWTH
(Asymptotic standard errors in parentheses)

Firms with 49 or fewer employees	$ln\ P_2/P_1$	$ln\ P_3/P_1$	$ln\ P_4/P_1$	$ln\ P_5/P_1$
Constant	1 88 (2 10)	-5 33 (2 77)	0 00737 (3,69)	-3 28 (4 62)
NE	0 875 (0 336)	0,824 (0 406)	-0 0308 (0 579)	0,338 (0 580)
NC	0 617 (0 322)	0.551 (0 385)	-0 633 (0.640)	0 211 (0 576)
South	0 555 (0 322)	0 800 (0 385)	-0,0503 (0 550)	0 178 (0 557)
Manufacturing	-1 21 (0 552)	-0 207 (0 679)	0 184 (0 898)	-0 182 (1,04)
Trade	-0 418 (0 460)	0 196 (0 603)	-0 270 (0 823)	0,177 (0 913)
Service	-0 249 (0 485)	0 154 (0 639)	-0 932 (0 903)	0 416 (0 934)
10-49 Employees	-0 892 (0 240)	-0 959 (0,439)	-0 800 (0 452)	0 702 (0 298)
% Change in Sales	0 0313 (0 00644)	0 0926 (0 0102)	0 0900 (0 0106)	0 592 (0 00778)
Knew of ETC	0 0926 (0 243)	0 666 (0 282)	0 669 (0 433)	0 582 (0 419)
Degrees of Freedom - 640		Log Likelihood = -714 6		
Firms with over 50 employees				
Constant	-0 440 (2 05)	-1 57 (1 84)	5 72 (5 55)	-9 07 (4 99)
NE	0 193 (0 369)	-0 205 (0 332)	-1 30 (0 917)	-0 423 (0 883)
NC	-0 175 (0 376)	-0 112 (0 322)	-0 751 (0 757)	-0,0540 (0 829)
South	-0 284 (0 370)	-0 307 (0 321)	-0 652 (0 723)	-0 213 (0 772)
Manufacturing	0 0332 (0 351)	0,612 (0 317)	1 60 (1 14)	0 656 (0 946)
Trade	0 184 (0 345)	0 421 (0 163)	2 04 (1 20)	0 752 (1,08)
Service	-0 200 (0 390)	0 280 (0 372)	0 105 (1 50)	1 46 (0 970)
250-499 Employees	0 0532 (0 283)	0 136 (0 249)	-1 18 (0 835)	-0 537 (0 704)
Over 500 Employees	0 0567 (0 298)	0 0912 (0 260)	-0 768 (0 667)	-0 777 (0 667)
% Change in Sales	0 0364 (0 00842)	0,0981 (0 00870)	0 144 (0 0174)	0 160 (0,0174)
Knew of ETC	0 0231 (0 277)	0 113 (0 249)	1 01 (0 621)	1,57 (0 833)
Degrees of Freedom = 694		Log Likelihood = -708 8		

[6] The results may not be invariant to the choice of the intervals. However, some experimentation leads us to conclude that the results are not highly sensitive to this choice. We tried 2, 3, 4, and 5 intervals and placed the divisions at different points. In no case were the results greatly changed.

ratio tests show that the knowledge variable does matter in all four equations collectively. The conscious effort variable was not used because it is clearly endogenous. The only variable which is asymptotically statistically significant in all the equations is the percent change of sales. The positive signs on its coefficients indicate that the higher is the percent increase in sales, the more weight there is in intervals S_2 through S_5 (i.e., increase in sales shifts the distribution to the right).

Since it is difficult to interpret the *log* odds results, we have unscrambled the *log* odds to obtain estimated probabilities. Due to the nonlinearity of the estimates, the probabilities must be evaluated at particular points. For example, consider a southern firm in a trade industry with the overall average increase in sales (12 percent). According to our estimates, 15.2 percent of such firms with between 10 and 49 employees have employment changes in the S_1 interval; 68.2 percent of the firms have employment growth in the S_2 range; and 0.8 percent, 14.7 percent, and 1.0 percent have growth in the S_3 through S_5 intervals, respectively. In other words, almost seven out of ten such firms have virtually the same employment in 1977 as in 1976, while 16.5 percent have at least 2 percent or more growth. The respective probabilities for firms which knew about *ETC* are 12.5, 61.3, 1.3, 23.6, and 1.4 percent. In other words, firms which knew of the credit have a distribution which is skewed to the right. Of those firms which did not know about the credit, 83.4 percent have no more than 2 percent employment growth (intervals S_1 and S_2); while only 73.8 percent of the firms which knew of the credit have 2 percent or less growth. Most of the shift in the distribution occurs out in the right tail; 15.7 percent of the ignorant firms have over 30 percent employment growth (S_4 and S_5) compared to 25.0 percent of the knowledgeable firms.

Firms with over 500 employees which did not know about the credit have probabilities of 36.4, 35.5, 26.8, 1.3, and 0.0 percent. Those which knew about the *ETC* have probabilities of 34.3, 34.2, 28.3, 3.2, and 0.1 percent. In other words, 71.9 percent of the firms which did not know about the credit have no more than 2 percent employment growth (S_1 and S_2), compared to 68.5 percent of the knowledgeable firms. Similarly, 5 percent more of the knowledgeable firms have greater than 2 percent employment growth (S_3, S_4 and S_5).

Knowledge of the *ETC* shifts more weight to the right tail of the employment growth distribution for small and moderate size firms than for large firms. For example, 9.5 percent more knowledgeable firms with 0–9 employees have more than 2 percent employment growth (S_3, S_4, and S_5) compared to 9.7 percent for firms with 10–49 employees, 5.0 percent for firms with 50–249 workers, 3.1 percent for firms in the 250–499 range, and 3.5 percent for firms with over 500 employees The size differences are even more pronounced in the extreme right tail of the distribution (S_4 and S_5).

These multinomial logit estimates indicate that firms which knew about the *ETC* are more likely to have large increases in employment in 1977 than other firms. The amount of the increase in employment from knowledge of the program decreases with firm size. This result is not surprising, since a firm with 2 employees which adds another will increase by 50 percent, while a firm of 500 employees which adds 47 more workers will not even increase 10 percent. Moreover, if the latter firm adds more workers it will not receive any marginal wage subsidy due to the $100 thousand limit.

IV. Conclusions

Our preliminary results indicate that the New Jobs Tax Credit may have shifted the distribution of the rate of growth of employment. Linear regression results indicate that those firms which knew of the program hired over 3 percent more workers than other firms. The multinomial logit study indicates that the *ETC* shifted the entire distribution of employment growth to the right. For example, for southern small firms in trade industries, over 9 percent more of the firms that knew about the *ETC* had employment growth over 2 percent, compared to similar small firms that did not know of the program. Over 3 percent more of the large firms who knew of the program had more than 2 percent employ-

ment growth compared to similar large firms who did not know of the program. A major problem in interpreting these results, however, is that knowledge of the program is not a random variable. One might regard these results as the upper bound on the short-run impact of this $1.7 million credit.

A simple examination of the descriptive statistics indicates that even if the tax credit substantially affected some firms, most firms either did not know of the program or were not influenced by it; a result which makes this short-run program an imperfect countercyclical tool. In addition based on theoretical considerations mentioned above, we believe that traditional monetary and fiscal policies are better suited to dealing with cyclical problems. On the other hand, the estimated impact of the current *ETC* among knowledgeable firms suggest that a permanent program aimed at increasing the *equilibrium* employment rate may be practical.

REFERENCES

O. Ashenfelter, "Evaluating the Effects of the Employment Tax Credit," in *Conference Report on Evaluating the 1977 Economic Stimulus Package*, U.S. Department of Labor, Office of the Assistant Secretary for Policy, Evaluation and Research, 1978.

R. Barro and H. Grossman, "A General Disequilibrium Model," *Amer. Econ. Rev.*, Mar. 1971, *61*, 82–93.

R. Eisner, "Tax Credits to Increase Employment and Reduce Inflation," in *Hearings on Administration of the Existing Tax Credit, Credit and Policy Alternatives for the Future*, Senate Subcommittee on Administration of Internal Revenue Code, Washington, July 18, 1978.

G. C. Fethke and S. H. Williamson, *Employment Tax Credits as a Fiscal Policy Tool*, a study prepared for the use of the Joint Economic Committee, 94th Cong., 2d sess., 1976.

J. M. Perloff, "Approaches to Evaluating the Local Public Works Program," in *Conference Report on Evaluating the 1977 Economic Stimulus Package*, U.S. Department of Labor, Office of the Assistant Secretary for Policy, Evaluation and Research, 1978.

M. L. Wachter, "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers*, Washington 1976, *1*, 115–59.

——, "Evaluating the 1977 Stimulus Package: A Summary Statement," in *Conference Report on Evaluating the 1977 Economic Stimulus Package*, U.S. Department of Labor, Office of the Assistant Secretary for Policy, Evaluation and Research, 1978.

U.S. Department of Commerce, "New Jobs Tax Credit Survey, Covering 1977 Tax Year," mimeo, Apr. 1978.

Congressional Budget Office, "Employment Subsidies and Employment Tax Credits," background paper, Apr. 1977.

Stimulating the Macro Economy Through State and Local Governments

By EDWARD M. GRAMLICH*

The economic stimulus program of early 1977 featured a strong dose of what might be termed indirect countercyclical policy. Rather than altering federal expenditures and taxes directly, the stimulus program consisted mainly of three different grant programs for state and local governments: a) countercyclical revenue sharing (*CRS*); b) public service employment (*PSE*); c) local public works (*LPW*). In the parlance of the public finance literature, the first of these grants was an unconditional block grant, the second was a close-ended categorical grant with no local matching for the purpose of stimulating local government employment, and the third was a close-ended categorical grant with no local matching for the purpose of stimulating local government construction.

Stimulating, or attempting to stimulate, aggregate spending through state and local governments in this way is fiscal federalism with a vengeance. The federal government is not abdicating its stabilization responsibilities and leaving it up to states and localities to do the job,[1] but it is placing its own stabilization policy at the mercy of the behavior of state and local governments. There are no restrictions at all on the use of the *CRS* grants—they can be spent, used for tax reduction, or used to rebuild financial net worth (asset stocks less outstanding debt), with only the first two uses having any stimulative effect at all. There are restrictions on what can be done with the other two grants, but the well-known displacement phenomenon implies that with these grants it also may be possible for states and localities to frustrate the restrictions and use the grants as they would any other source of revenue sharing. What happens to all three grants then is an empirical issue, and the timing and magnitude of any stabilization impact depends on how the numbers come out.

In this paper I briefly describe a model for estimating this stabilization impact and show what it suggests for the three grants. Only the *PSE* grant will be seen to have any positive short-run impact on aggregate spending at all, and that impact will prove to be both diluted and short-lived, indicating that the general idea of stimulating the economy through state and local governments is probably not a very good one. Plain old permanent federal income tax cuts retain their superiority as a fiscal stabilization device. But even though as *stimulation devices* the three grant programs leave much to be desired, as *policies* they may still be valuable, and the paper also suggests how one might do a more complete evaluation of each of the grant programs.

I. The Empirical Model

On three previous occasions I have tried to estimate a quarterly time-series model of the behavior of state and local governments in the national income accounts (*NIA*). Here I repeat and update the procedure, but I do not describe it. The underlying conception of each effort was to use consumer utility-maximization theory to show how an aggregation of state and local governments would respond to changes in community disposable income, relative prices, demographic changes, interest rates, stocks of assets, and federal grants of various types by altering current and capital expenditures, taxes, and the budget surplus. The estimates have been made subject to three separate accounting or economic constraints:

a) *The Adding Up Constraint:* Any variable that directly enters the state and

*University of Michigan. I would like to thank Michael Wolkoff for doing most of the computer work, and Laurie Bassi, Robert Cline, Alan Fechter, Daniel Hamermesh, David Levin, and John Palmer for their comments on an earlier draft.

[1]An abdication likely to result in no stabilization action being taken, see Wallace Oates, Appendix to ch. 1

local budget (such as federal grants) must be exactly allocated to all other uses of funds; while any variable that does not directly enter the budget must have effects that are offset elsewhere in the budget.

b) *Stock Adjustment:* For both physical capital and financial asset stocks, utility is derived from the stock itself, not the budgetary flow. In the long run the model behaves as a stock adjustment model in this regard, with both net investment and net financial saving being only temporarily altered in response to some change in an exogenous variable.

c) *Grant Distinction:* Since an important use for the model is to distinguish the effects of different types of grants, grants are treated differently according to their restrictions. Open-ended price reduction grants are viewed as altering relative prices, unconditional block grants are viewed as shifting out the budget constraint line, and close-ended categorical grants are viewed as moving

TABLE 1—CONSTRAINED ESTIMATES OF THE MODEL

Independent Variables	Dependent Variables E_1	E_2	E_3	$-T$	$F_{-1} + S$
$F_{-1} + X$		.0327		.0580	.9093
		(3.5)		(4.8)	(46.3)
$.67Y + .33Y_{-1}$	.0269	.0150	.0287	−.0922	.0216
	(4 2)	(2.0)	(2 7)	(−10.0)	(–)
$.25\sum_{j=0}^{3} PSE_{-j}$	−1.0690				1.0690
	(−3.9)				(3.9)
$(1/m_1)G_1$	− 9356				.9356
	(−25.7)				(25.7)
$(1/m_2)G_2$		− 9453			.9453
		(−22.3)			(22.3)
K_{-1}	.0261	.0306	− 0202		−.0365
	(8.1)	(7.1)	(−3.0)		(–)
W	−150.6				150.6
	(−5.1)				(5 1)
LPW			−20.92		20.92
			(−3 7)		(3.7)
FEM	−3.964	8.228			−4.264
	(−1 6)	(2 4)			(–)
UR	.9692	1.600			−2.5692
	(2 2)	(2.8)			(–)
$\bar{R}^2$ (diff)	.92	.88	.05	.28	95

Notes: The *NIA* government accounts, eliminating all social insurance trust fund items, are the basic data set.

Definitions: Taxes T equal all taxes plus surplus of government enterprises, Discretionary spending for wages E_1 equals total wage bill payments less PSE grants less mandated wage expenditures on other categorical grants $(1/m_1)G_1$, where m_i is the federal share; Discretionary spending for other current purchases and transfers E_2 equals total spending for these purposes less grant mandated expenditures $(1/m_2)G_2$; Discretionary spending for construction E_3 equals total spending less grant mandated expenditures $(1/m_3)G_3$; Exogenous budgetary inflows X equals general revenue sharing GRS plus countercyclical revenue sharing CRS less interest and debt service payments D less mandated expenditures on all federal categorical grants $\Sigma_{i=1}^{3}((1/m_i) - 1)G_i$; Financial surplus, or budget surplus S, equals $X + T - \Sigma_{i=1}^{3} E_i$; Financial stocks F equal $S + F_{-1}$; Capital stocks K equal $E_3 + (1/m_3)G_3 + (1 - \delta)K_{t-1}$, where $\delta = 005$ is the quarterly depreciation rate; Local public works LPW equal a dummy variable building up from 1976II to 1976IV and then remaining at 1 0 through 1977IV; Income Y equals GNP less federal taxes; Wage rates W equal an index of the average compensation rate for state and local employees (1972 = 1.0); Demographic terms are the proportion of families headed by females FEM and the constant demographic weight unemployment rate UR; Total state and local expenditures EXP equals $\Sigma_{i=1}^{3}(E_i + (1/m_i)G_i) + PSE$; PSE grants are measured from PSE employment, the most reliable figure. To put the variable in real dollars, employment is multiplied by the constant 1972 annual wage of public employment workers ($8,200)

All variables are estimated as first differences of the variable in real per capita terms (except W, which is the first difference of the relative wage and the demographic terms which are simple first differences). t-ratios below coefficients, (–) if not calculated

TABLE 2—SHORT- AND LONG-RUN RESPONSE OF *EXP*, *T*, AND *S* TO CHANGES IN *CRS* AND *PSE*[a]

Time Passed	Deviations from Initial Values When the Federal Government: Raises *CRS* by 1.0				Raises *PSE* by 1.0			
	EXP	*T*	*S*	*F* − 1	*EXP*	*T*	*S*	*F* − 1
1	.033	−.058	.909	-	.738	-	.267	-
2	.063	−.111	.826	.909	.480	−.015	.505	.267
3	.090	−.159	.751	1.735	.229	−.045	.726	.772
4	.115	−.202	.683	2.486	−.014	−.086	.928	1.498
5	.138	−.242	.620	3.169	.016	−.141	.843	2.426
6	.158	−.278	.564	3.789	.044	−.190	.766	3.269
7	.177	− 310	.513	4.353	.069	−.234	.697	4.035
8	.194	−.340	.466	4.866	.092	−.274	.634	4.732
9	.209	−.367	.424	5.332	.112	−.311	.577	5.366
∞	.362	−.638	-	10.025	.320	−.680	-	11.720

[a]All variables are defined in notes to Table 1.

governments to the kink point in the constraint line.

The results of estimating this model quarterly from 1954 through 1977 are given in Tables 1 and 2. Table 1 presents coefficient estimates and fit statistics for a constrained estimate of the model. Blanks in the table indicate cases where the variable was not statistically significant with the proper sign, and was therefore constrained to equal zero in the final refitting. Table 2 then gives the dynamic implications of changes in *CRS* and *PSE*. As a general matter, both the coefficient estimates and the dynamic patterns in the updated version of the model are quite similar to what they were in earlier incarnations.[2]

II. The Programs

The first program examined with this model is *CRS*. The estimates of Tables 1 and 2 imply that in the short run only \$.03 of a dollar's revenue sharing grant will end up as expenditures and \$.06 as tax reduction, but that in the long run \$.36 will go into expenditures and \$.64 into tax reduction. The reason for this behavior is that in the short run a large share of the grant pads surpluses, but as financial stocks cumulate there is progressively less reason for governments to save and they gradually raise spending and cut taxes until the impact on the surplus is nil.

These results as they stand imply that not much of a macro-stimulation case can be made for *CRS*: the money is spent only very slowly, and much of it is simply tax reduction which wouldn't be expected to have effects much different from the more prompt direct federal tax cuts. But such an assertion has two possible drawbacks: one econometric and one philosophical. The econometric one is that the coefficients are derived from those on *X*, budgetary inflows of all sorts. There is nothing wrong with making such an inference in theory, for *CRS* is precisely an unconstrained budgetary inflow just like every other positive or negative component of *X*. There is also little else one could do in practice, at least in a time-series context, for *CRS* grants have existed only in the last year of the estimation period (for what it's worth, the nonconstruction residuals were very small then). But it may still be risky to make such an inference. Under the present law, *CRS* exists only in high unemployment years (the overall rate must exceed 6 percent), and is paid only to governments of areas experiencing high unemployment (in excess of 4.5 percent). Hence there is a greater likelihood that *CRS* funds will be used for maintaining programs that would otherwise have to be killed in a cyclical downturn, or for preventing tax rate increases, than there is for the other components of *X*. If such is the case, the macro-stimulation benefits of *CRS* will be greater

[2]The most thorough description was in my 1973 paper with Harvey Galper, and the most recent in my 1978 paper.

than those noted in Table 2.

There is a more basic point: *CRS* is a cyclical program, and as such only one of its possible benefits is as an automatic aggregate demand stabilizer. The other conceivable benefit is as a form of economic disaster insurance for state and local governments. More and more these governments rely on cyclically sensitive income and sales taxes, and indeed even property taxes could be somewhat cyclically sensitive with up-to-date reassessments. On the expenditure side, the growth of unions, wage contracts, and tenure arrangements implies that wage expenditures are becoming more difficult to alter in the short run, and the growth of public assistance transfers indicates these expenditures may be also. Hence it could be argued that state and local governments are now quite vulnerable to the business cycle and need a form of disaster insurance to prevent costly interruptions of services in a downturn. Whether this argument is at all convincing depends on whether various state and local governments do save for cyclical exigencies, whether this saving will be reduced by a federal cushion, and whether the politics of *CRS* enables cyclical funds to go where they are most needed. Each of these is a complex question that cannot be dealt with here, but what can be said is that looking at macro stimulation is only half of the story. If the macro-stimulation benefits of *CRS* are nil, it is somewhat harder to justify the program, but by no means impossible.

The next program is *PSE*. According to the estimates of Tables 1 and 2, the so-called displacement effect is very strong, leading to no impact of *PSE* on total expenditures after four quarters.[3] But if grant displacement is strong, state and local governments must necessarily experience a rise in their surplus, and this addition to financial stocks then encourages spending and tax reduction as with revenue sharing. Hence after hitting this nadir after four quarters, the fiscal impact of *PSE* then begins rising—following a path approximately like delayed-reaction revenue sharing.

Again the problems in believing these results too religiously can be grouped into the statistical and the philosophical. On the statistical side, *PSE* has not been lumped with any other programs the way *CRS* was, but it has changed in character over time. In particular, the program was more tightly constrained to try to insure employment increases by the Carter Administration in early 1977 (for what it's worth, the wage bill residuals are not positive in the last two quarters of 1977). Of perhaps more importance is the fact that an estimated 26 percent of the Comprehensive Employment Training Act (*CETA*) money simply passes through local governments on the way to private nonprofit agencies known as community based organizations. Due to a soon-to-be-remedied accounting mistake, the *NIA* includes this money in the *PSE* grants but not anywhere in budget expenditures, implying that the displacement effect will inevitably be overstated.[4]

A related problem refers to the choice of the dependent variable. The equation in Table 1 uses the real wage bill from the *NIA*, a direct component of real *GNP*. But it would also be possible to use the employment of state and local governments from establishment employment data. Were this to be done, the coefficient of *PSE* would rise from −1.07 in the wage bill variant to −.60 in the employment variant, indicating only 60 percent displacement.[5] This suggests either

[3]Essentially, the timing and magnitude found by George Johnson and James Tomola. The lag structure used here is obviously arbitrary, but unlike Michael Borus and Daniel Hamermesh, I found that various simple lag polynomials gave about the same steady-state results

[4]The 26 percent number comes from a recent Urban Institute telephone survey conducted by Laurie Bassi and Alan Fechter. The *NIA* "mistake" was candidly admitted by Bureau of Economic Analysis (*BEA*) personnel, and was corrected in the revisions of 1978. Those numbers were released too late for me to use in this paper.

[5]The equation can only be estimated over the 1967–77 period because the establishment data only go back that far. It is

$$SL\ Emp - PSE\ Emp = \underset{(-1.0)}{-.596}\ PSE\ Emp + \underset{(0.08)}{.0016}\ (.67Y + .33Y_{-1}) + \underset{(3.5)}{.0033}\ K_{-1} - \underset{(-0.7)}{7.46}\ W - \underset{(-0.1)}{.126}\ FEM + \underset{(0.5)}{0.1}\ UR,\ \bar{R}^2 = .07$$

G_1 is omitted from both sides of the equation because

that there may be some full-time/part-time shift (in which case the *NIA* wage bill estimate is the better one) or that *PSE* may lower the average wage paid state and local employees to a degree unaccounted for by *NIA* deflation. If the latter is the case, the best displacement estimate is that from the employment numbers, and the lower wage may also be considered a benefit of the program. If the unmeasured output of public sector workers remains constant, the lower *PSE* wages in effect generate a transfer away from public sector workers and could perhaps even make *PSE* a weapon in the government's fight against wage inflation. Before that is loudly praised, however, more effort should be expended on data collection to make sure the *NIA* can properly measure this reduction in average public sector wage rates.[6]

Finally let us turn to the basic point. The *PSE* is in part a stabilization program and in part a program aimed at improving the competitive lot of disadvantaged low-wage workers. Suppose there is 100 percent displacement. If the *PSE* wage ceilings are enforced, the program presumably is stimulating relatively low-wage *PSE* employment and reducing higher-wage regular employment. This transfer of employment demands and workers' producer surplus goes from high-wage regular employees to low-wage or underemployed *PSE* employees, and again could be very desirable from a social standpoint. As with *CRS*, macro stimulation is not everything, and the degree to which high displacement is used as evidence against *PSE* may be quite excessive.

The final grant program is *LPW*. This grant was sufficiently unique that I did not even try to incorporate the variable into the regular model, but simply used a dummy variable. The reader may be surprised to find a negative coefficient: how can a grant to stimulate local construction actually reduce it? The answer, given in more detail and (some say) more melodramatically in my 1978 paper, can be found in a careful examination of the details of the bill. This bill gave free (no-match) money to state and local governments for construction projects that could be started within 90 days, with the intragovernment allocation of funds to be decided administratively. The Economic Development Agency was flooded with applications totalling $24 billion for the initial $2 billion of funds, and the unfunded governments were not told to go back and now build their projects but encouraged to wait until next year (1977) when another $4 billion would be forthcoming. In such circumstances it became quite rational for governments to hold up projects that would have otherwise been started to see if federal funds were to be forthcoming, and quite possible for *LPW* to have a negative short-run effect on construction. The actual estimated reduction of Table 1 of $6 billion in nominal terms (multiplying 20.92 by the price level and population) is indeed moderate both in relation to the queue of unfunded projects and the otherwise mysterious drop in state and local real construction in 1976–77.

As macro stabilizers, then, none of the grant programs come out very well. The *CRS* seems to have effects that are very small in the short run, *PSE* to have effects that are very transitory, and *LPW* to have effects that are perverse. One lesson is that an economic stimulus program should not rely only on state and local grants; some other means must be found to stabilize the national economy. But a second lesson that would appear to follow does not. It is not obvious that *CRS* and *PSE* should be scrapped even though as macro stimulants they are ineffective. In both cases a full evaluation should delve into more subtle considerations of the sort mentioned, but not dealt with, here. Conceivably one could even make a similar case for *LPW*, but there the macro-stimulation impact is so perverse and the other benefits so dubious that the evaluation reasoning would have to be very subtle indeed.

---

there are no data on employees hired under categorical grant programs, and the dependent and first independent variable are in per capita difference form. The fit statistics are all substantially worse than with the wage bill equation, and the coefficients much less reliable, but they do indicate less displacement.

[6]These comments are not meant to be critical of *BEA*. Already it does try to deflate *PSE* and non-*PSE* employment separately, and its only difficulty in doing that is that there are not good data on average wages for *PSE* and regular employment.

REFERENCES

M. E. **Borus and D. S. Hamermesh,** "Study of the Net Employment Effects of Public Service Employment—Econometric Analysis," mimeo, National Commission on Manpower Policy, 1978.

E. M. **Gramlich,** "State and Local Budgets the Day After It Rained: Why is the Surplus So High?," *Brookings Papers*, Washington 1978, *1*, 191–214.

—— **and H. Galper,** "State and Local Fiscal Behavior and Federal Grant Policy," *Brookings Papers*, Washington 1973, *1*, 15–58.

G. E. Johnson, and J. D. Tomola, "The Fiscal Substitution Effect of Alternative Approaches to Public Service Employment Policy," *J. Hum. Resources*, Winter 1977, *12*, 3–26.

W. E. Oates, *Fiscal Federalism*, New York 1972.

Economic Development and the Theory of International Trade

By RONALD FINDLAY*

Recently the major issue in the international aspects of economic development has been the so-called "North-South dialogue" in connection with the *UN* resolutions calling for a New International Economic Order. The intellectual basis for the proposed reforms, in so far as one exists, appears to lie in the well-known writings of Raul Prebisch and Hans Singer. Both of them argue that there is a fundamental asymmetry in the workings of the global economic system which biases the resulting income distribution in favor of the industrial North and against the predominantly primary producing South. Neither writer has been successful in putting forward convincing arguments for such asymmetry. The standard trade theory of the Heckscher-Ohlin variety is usually presented in such a way that "countries *A* and *B*" are identical in all respects except for a difference in factor proportions that leads to pretrade product and factor-price differentials that are removed by free trade. There is no room for any asymmetry here. It would therefore seem to be both relevant and interesting to construct and investigate models that exhibit the Prebisch-Singer asymmetry at the level of rigor that generally prevails in pure trade theory. The rest of this paper will present two examples of such models from current research. The first consists of a simple diagrammatic exposition of an interesting but heavily mathematical paper by Murray Kemp and M. Ohyama and the second outlines the essential features of an approach to the analysis of North-South economic relations found in my earlier paper.

*Columbia University.

I. The Kemp-Ohyama Model

The structure of the world economy assumed by Kemp-Ohyama is one in which the North and South each have a fixed endowment of labor and of capital, which I denote $\mathbf{K}_N$ and $\mathbf{K}_S$. Labor is tied to each region but capital is internationally mobile in response to any differences in return. The capital used in production in each region, K_N and K_S, must satisfy the relation $K_N + K_S = \mathbf{K}_N + \mathbf{K}_S$. The South produces a raw material R by means of labor and capital according to the production function $R(K_S)$, where the fixed labor input is suppressed, with a positive first and negative second derivative with respect to K_S. The North produces a final product F with its fixed labor force and with R and K_N as the variable inputs according to the production function $F(K_N, R)$, with positive first and negative second derivatives with respect to each argument and a nonnegative cross derivative.

Suppose, to begin with, that there is complete free trade. A key relationship can then be readily derived which is

$$\frac{\partial F}{\partial K_N} = \frac{\partial F}{\partial R} R'(K_S) \tag{1}$$

The logic of this relation is easy to see. Owners of capital in either region can get $\partial F/\partial K_N$ by using their capital in the North or they can get $R'(K_S)$ of the raw material in the South and then transform this amount into $\partial F/\partial R$ of the final product. The "direct" use of capital $\partial F/\partial K_N$ must therefore be equally profitable at the margin with the "indirect" use $(\partial F/\partial R)\, R'(K_S)$. Dividing both sides of (1) by $\partial F/\partial R$ shows that the marginal rate of substitution between capital and the raw

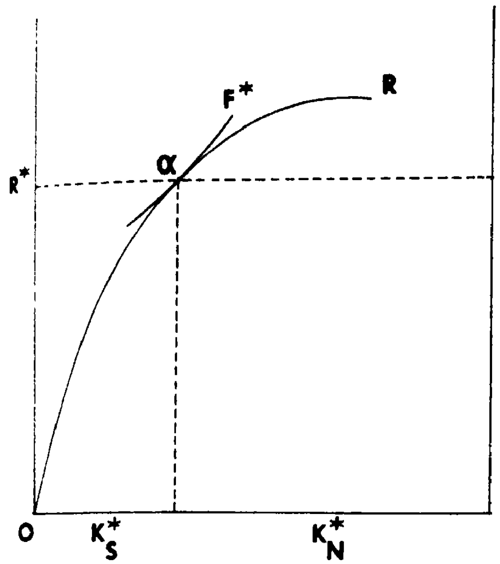


FIGURE 1

material in making the final product must be equal to the marginal productivity of capital in making the raw material.

Figure 1 gives a complete picture of the competitive solution. The horizontal axis measures the total capital in the world economy $\mathbf{K}_N + \mathbf{K}_S$ with K_S measured from the left and K_N from the right. The curve OR is the production function $R(K_S)$. The relation (1) is satisfied at the point α where OR is tangential to an isoquant representing the highest output of the final product attainable with the given amount of world capital, denoted F^*. Also determined by α are the corresponding equilibrium values R^*, K_N^*, and K_S^*. The ratio of the price of the raw material to that of the final product, which defines the terms of trade, will be equal to $\partial F(K_N^*, R^*)/\partial R$. The value of exports from the South in terms of the final product is $(\partial F/\partial R)R^*$. The direction of capital movement depends upon whether $K_S^* \gtrless \mathbf{K}_S$ and the value of interest payments and therefore of the export surplus is $(\partial F/\partial K_N)\ (K_S^* - \mathbf{K}_S)$. It is clear that the ownership distribution of the fixed world capital is irrelevant to the determination of F^*, R^*, K_N^*, and K_S^*. World income and consumption are equal to F^*, the distribution between the two regions depending upon the terms of trade $\partial F/\partial R$, the return on capital $\partial F/\partial K_N$ and on $\mathbf{K}_N$ and $\mathbf{K}_S$.

Suppose now that the North imposes an ad valorem tax on the export of the final product. It is easy to see that this will have no effect on northern capitalists since they can still obtain $\partial F/\partial K_N$ if they invest at home and $(\partial F/\partial R)R'(K_S)$ if they invest abroad, import the raw material obtained and use it in domestic production. The key relation (1) therefore continues to hold for northern capitalists in spite of the export tax in the North. Southern capitalists who invest in the North will have their return on capital reduced by the export tax that they have to pay when they repatriate their real earnings but they will lose exactly the same amount if they use their capital at home to produce the raw material, transfer it to the North to produce final output, and repatriate their real earnings. The export tax in the North reduces both sides of (1) by the same amount and therefore does not affect the incentives of southern capitalists as to the allocation of their capital since they suffer the same impoverishment in any event. Consequently K_S^* and K_N^* and therefore R^*, F^*, and $\partial F(K_N^*, R^*)/\partial R$ are not affected by the export tax in the North. Since $\partial F/\partial R$ is equal to the price of the raw material in the North the fact that it is left unchanged must imply that the full incidence of the export tax levied by the North falls on the South. If e denotes the ad valorem export tax rate, the terms of trade of the South must deteriorate to $1/(1+e)\ \partial F(K_N^*, R^*)/\partial R$ independently of how high e is. There is therefore no limit to the extent to which the North can "exploit" the South by the imposition of an export tax. This is the central striking result of the Kemp-Ohyama paper. It can also readily be seen that the same results as an export tax could always be achieved by means of a tax on the import of the raw material combined with a subsidy on the export of capital to the South to offset the impact of the import tax on the allocation of capital.

Further insight into the reason for the Kemp-Ohyama result is provided by an alternative interpretation of (1) that follows if both sides are divided by $R'(K_S)$. The right-

hand side is now just $\partial F/\partial R$ which is a decreasing function of R and an increasing function of K_N. With total world capital fixed, however, R can only increase if K_S increases and therefore if K_N decreases. Since the cross derivative in the North's production function is nonnegative, an increase in R must therefore lead to a decrease in $\partial F/\partial R$. Now the price the South receives for the production and export of a given level of R is precisely $\partial F/\partial R$, so that $\partial F/\partial R$ as a function of R is the demand curve for the South's exports. The left-hand side of (1) after division by $R'(K_S)$ is $(1/R'(K_S))$ $(\partial F/\partial K_N)$. The reciprocal of $R'(K_S)$ is the marginal cost of R in terms of capital, while $\partial F/\partial K_N$ is the opportunity cost of this capital in terms of final product that could be obtained by investment in the North, so that the product of the two terms is the marginal cost of the raw material in terms of the final good. Increasing R is accompanied by falling $R'(K_S)$ and rising $\partial F/\partial K_N$ so that the marginal cost rises as R increases. The competitive equilibrium output R^* is where the demand curve cuts the marginal cost curve or where price equals marginal cost, which is the alternative interpretation of (1). The export tax imposed by the North reduces the price and the marginal cost of R in the same proportion, leaving R^* unchanged, and the full incidence of the export tax falls on the terms of trade received by the South.

This interpretation also enables one to examine the optimal policy for the South on the assumption that the North is passive. The competitive output R^* is where marginal cost equals price. The optimal monopoly policy would be to restrict output and raise price to the point at which marginal cost equals marginal revenue instead of price. This is nothing but the familiar optimal tariff argument. The salient points to note in the present context are that the optimal policy for the South involves reducing the output of the final good and therefore world consumption and also that the extent to which the South can hurt the North is limited by the elasticity of the North's demand curve for the raw material. The Kemp-Ohyama model thus provides a simple way in which to formalize the "unequal bargaining power" between exporters of raw materials and of manufactures that Third World spokesmen have complained about bitterly for a long time. It should be pointed out, however, that the analysis is confined to taxes and subsidies and does not extend to physical measures such as an embargo.

II. The Terms of Trade and Steady-State Growth in an Asymmetrical World Economy

This section sketches another stylized model of North-South economic relations, this time in the context of growth. Dynamic models of growth and trade have proliferated in recent years, the one by H. Oniki and Hirofumi Uzawa being perhaps the most influential. It is not, however, a very appropriate tool for the analysis of North-South relations since the two regions are identical in all respects except for a difference in the savings ratio and full employment is always assumed to prevail in both regions. Also the stability condition of the capital-goods sector being more labor intensive imposes the anomalous outcome of the less developed region being the capital goods exporter. In the model to be discussed here there is an asymmetry in the labor market between the two regions, the North having full employment with a flexible real wage and the South a fixed real wage and variable employment. The North is completely specialized in a single composite manufactured good, which can be used for either consumption or investment as in the Robert Solow neoclassical growth model. A constant fraction of income s is saved and there is a fixed growth rate n of the labor force. Indeed, the North's economy is identical with the Solow model except for the fact that a proportion of consumption expenditure, depending on relative prices, is spent on the import of a composite primary commodity from the South, which completely specializes on its production. The output of both manufactures and the primary commodity are subject to neoclassical production functions $f(k_N)$ and $\pi(k_S)$ with constant returns to scale and labor and capital as inputs, k_N and k_S denoting the capital-labor ratios. Capital consists of a stock of manufac-

tures in each case, which can be augmented by investment. Depreciation is ignored for simplicity. Saving in the South is a constant fraction σ of profits, all wages being consumed. The South also consumes both products with demand depending on relative prices. Income elasticities of demand are assumed to be unity in both regions.

The South is a dual or labor surplus economy à la W. Arthur Lewis, with a perfectly elastic supply of labor to the primary export sector at a fixed real wage in terms of the primary commodity. The source of the labor supply is a subsistence sector or "peasant hinterland," which is outside the model except for its role as a reservoir of labor to the primary export sector. Profit maximization will make the marginal product of labor equal to the fixed real wage and this determines the capital-labor ratio k_S^* in primary production, the output per unit of labor $\pi^* = \pi(k_S^*)$ and the marginal product of capital $\pi'(k_S^*)$. The rate of profit will be $\theta\pi'(k_S^*)$ where θ is the ratio of the price of the primary commodity to the price of manufactures—the terms of trade. The rate of growth of capital, output, and employment in primary production will be $\sigma\theta\pi'(k_S^*)$ by virtue of what Joan Robinson calls the Anglo-Italian equation. In the steady state the growth rate of the North must be equal to its exogenously given natural growth rate n For a steady state in the world economy as a whole the growth rates of both regions must be equal which implies that the only value of the terms of trade θ consistent with long-run equilibrium is

$$\theta^* = \frac{n}{\sigma\pi'(k_S^*)} \tag{2}$$

a remarkably simple expression, independent of the production function and saving ratio of the North and of consumer preferences in both regions.

The unique steady-state capital-labor ratio in the North k_N^* is determined by the requirement that $sf(k_N^*) = nk_N$ and this in turn determines the per capita output in the steady state as $f^* = f(k_N^*)$. With per capita income in the North and the terms of trade set as f^* and θ^*, respectively, the import propensity of the North is determined at $m^* = m(\theta^*)$ and the per capita expenditure on imports from the South is m^*f^*. Similarly π^* and θ^* determine μ^*, the propensity of the South to import manufactures for investment as well as consumption. Thus the per capita demand for manufactured imports is $\theta^*\mu^*\pi^*$. If $L_N(t)$ and $L_S(t)$ denote employment levels in the two regions at any time t then the requirement of balanced trade imposes the condition

$$L_N(t)\, m^*f^* = L_S(t)\theta^*\mu^*\pi^* \tag{3}$$

from which it follows that

$$\lambda^* \equiv \frac{L_S(t)}{L_N(t)} = \frac{m^*f^*}{\theta^*\mu^*\pi^*} \tag{4}$$

The balanced trade condition therefore determines the scale of the export sector in the South relative to the economy of the North. The convergence of λ, k_N, and θ to their steady-state values is proved in my unpublished paper to which the reader is referred for further details and extensions.

The asymmetrical properties of the world economy in this model emerge clearly from the steady-state relations (2) and (4). The rate of growth of the world economy as a whole is determined exclusively by the natural growth rate of the North. The size of the North's economy at any instant is always fixed by the level of its fully employed labor force whereas the level of export employment in the South depends, in addition to its own parameters, on the propensities to save and import in the North since these determine m^* and f^* in (4). An increase in s leaves θ^* unchanged whereas an increase in σ reduces θ^* in the same proportion. Similarly a shift in the North's production function leaves θ^* unchanged while an increase in $\pi'(k_S^*)$ at a constant real wage reduces θ^* by the same proportion. The model therefore illustrates the Prebisch-Singer contention that the "fruits of progress" are preserved by the North in higher incomes while they are dissipated for the South in deteriorations of the terms of trade. However, the model shows that the South does benefit from greater saving and productivity in the North in the form of a relative expansion of scale while

increases in θ^* are accompanied by increases in λ^* as well.

The purpose of this paper has been to indicate that it is possible to construct rigorous models in which asymmetric features in the world economy, as stressed in the recent North-South dialogue, appear. This should not be taken, however, as in any way necessarily implying or endorsing the conclusions drawn and policy prescriptions recommended by those voices in this dialogue that stress this asymmetry.

REFERENCES

R. Findlay, "The Terms of Trade and Equilibrium Growth in the World Economy," disc. paper no. 77-7822, The Economics Workshops, Columbia Univ., Mar. 1978.

M. C. Kemp and M. Ohyama, "On the Sharing of Trade Gains by Resource-poor and Resource-rich Countries," *J. Int. Econ.*, Feb. 1978, *8*, 93–115.

W. A. Lewis, "Economic Development with Unlimited Supplies of Labor," *Manchester Sch. Econ. Soc. Stud.*, May 1954, *22*, 139–91.

H. Oniki and H. Uzawa, "Patterns of Trade and Investment in a Dynamic Model of International Trade," *Rev. Econ. Stud.*, Jan. 1965, *32*, 15–38.

Raul Prebisch, *The Economic Development of Latin America and its Principal Problems*, New York 1950.

H. W. Singer, "The Distribution of Gains between Borrowing and Investing Countries," *Amer. Econ. Rev. Proc.*, May 1950, *40*, 473–85.

R. M. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Feb 1956, *70*, 65–94.

Efficiency of *LDC* Trading Patterns: The Case of Iran

By HOSSEIN ASKARI, JOHN T. CUMMINGS, AND GUNTER RICHTER*

International trade theory has traditionally been cast in an idealized setting, with numerous assumptions, where the flow of goods is determined by static comparative advantage. In the Ricardo-Torrens theory, differences in technology across countries affect labor productivities, "pretrade" prices, and thus determine comparative advantage and trade. In the Heckscher-Ohlin theory, differences in relative factor endowments affect "pretrade" relative factor prices, relative costs of production, and thus determine comparative advantage and trade. Originally, such models were formulated for a two-country world, but with the addition of a third country it became clear that under certain circumstances multilateral comparisons may be appropriate (see Ronald Jones; Anne Krueger).

Empirical verifications of both theories have been quite extensive. Most of the tests have been bilateral, one country to another or one country trading with the rest of the world. In general the results have been, at best, very mixed. The explanations for the unexpected results have involved such theoretical modifications as the need to introduce natural resources explicitly, to distinguish between new and standard commodities, to account for heterogeneity of factors and products, to include barriers to trade and so on.

*Professor of international business; assistant professor of international business; and graduate student, Graduate School of Business, The University of Texas-Austin, respectively. Askari is grateful for support from a George Kozmetsky Fellowship, the Glenn and Betty Mortimer Excellence Fund, and the Center for Middle Eastern Studies. We are grateful to Charles Kindleberger for helpful comments, but remain responsible for any errors. Detailed tables and classifications are available on request. Much of the motivation for the paper is due to private discussions and interviews with a few Iranian businessmen. They in fact supplied two of the major reasons for our results; that of fragmentary knowledge of world markets and restrictive contracts of major multinational firms. It was because of their insight that we did not build on traditional theory, but asked a few simple questions.

In all this work, the basic existence of competitive markets has been an accepted assumption: that is, a country will face one import price for a homogeneous product because competition will drive out inefficient exporters. However this is a testable proposition; specifically, does a country buy its imports from the low cost producer and, if not, is the overpayment significant?

The a priori reasons for the existence of any such trade inefficiencies in import prices or price divergences are many. One possible explanation is that there exist variations in unit export prices for an exporter. Gary Hufbauer and J. P. O'Neill in examining unit values of *U.S.* machinery exports suggested several explanations for export price differences for a given exporter:

> Apart from qualitative differences, other causes have also been suggested for intercountry variance in export unit values. P. T. Knight raised the possibility of price discrimination between purchasers, while Bhagwati (1964, 1967) and Winston (1970) have suggested that tariffs and overvalued currencies could lead to fake invoices, and thus to apparent variations in export unit values. [pp. 265–66]

A second explanation would take issue with the implicit assumption that imports are channeled through one entity with perfect information. In reality imports are contracted for by many companies or individuals with fragmentary knowledge of world markets. Third, existing trading patterns are in large part determined by historical ties and are a function of past patterns of trade and past prices but may not accurately reflect current prices. Fourth, some current trade is based on past contracts and thus may differ from spot prices. Fifth, contracts may be made in any currency and thus fluctuations in exchange rate can affect the dollar price from one source vs. another. Sixth, quality differences

and especially quality control may be reflected in price. Seventh, quick delivery may require a higher price. Eighth, a large part of imports, especially in *LDCs*, is from multiproduct firms which place restrictions on their importing agents. Major importers in *LDCs*, contracted to multinational firms, are committed to import all the products of that multinational firm and cannot import a competitive product from the low-cost producer in the world. At the same time, other importers in that country may be insignificant or may have very little knowledge of world markets or barriers to entry may be significant. In a sense, international trade does not practically occur in single products but in groups of commodities that a company manufactures. Ninth, differential credit terms of contract or transport costs could contribute to variations in import prices. Finally, some countries, due to customs union arrangements, may pay above the world price for some goods.

Such imperfections outlined above should be most significant for *LDC*'s. Our country of interest is Iran. We examine Iran's import patterns from 1972 to 1974. Iran, although a member of Regional Cooperation for Development (with Pakistan and Turkey), is free from any significant customs union arrangement. In addition, given the rapid expansion of imports in 1973–74, one may note an interesting pattern of imports. We also compare Iran's experience to that of Sweden, a developed country, and India, an *LDC*, to determine if Iran is a special case.

The existence of large price discrepancies for homogeneous import products would undermine one of the assumptions of international trade theory. If such trade inefficiencies are significant, then it would give a strong reason for the mixed empirical results of trade theories. Furthermore, it would indicate to countries an important area of policy for reducing the size of their import bills.

I. Calculation and Data

For Iran we have made two basic sets of calculations. In the first we identified the lowest-cost producer in the world among major exports of that commodity. In Iran's case, a major exporter is defined as a country whose quantity of exports is sufficient to supply all of the import needs of Iran for a particular category. We then calculated what would have been Iran's import bill for that commodity had it been purchased from the lowest-cost producer. We then subtracted this hypothetical low import bill from the actual import bill to arrive at a measure of "overpayment" for a particular category. Second, we calculated this overpayment from a different perspective. Instead of using the lowest-cost producer as above, we found the low-cost producer from the countries that actually sold to Iran. These two sets of calculations were also done for Iran (1974) for imports from Organization for Economic Cooperation and Development (*OECD*) countries at the 4- and 5-digit SITC levels.

In order to compare the Iranian experience with another *LDC*, we did the same two sets of calculations for India (1974); India is a nation beset with economic problems and thus could little afford to overpay for its imports. In order to eliminate the possibility that such overpayments may be purely an *LDC* problem, we also did similar calculations for Sweden (1974). Finally, in order to check what happened on the export side (to corroborate Hufbauer and O'Neill), that is, countries charging different prices to different importers, we calculated the variance in export prices for India (1974) at the 3-digit SITC level. The 3-digit SITC data was taken from various issues of both *Yearbook of International Trade Statistics* and *Commodity Trade Statistics*. The 4- and 5-digit SITC data for *OECD* trade with Iran were taken from *OECD, Series C, Trade by Commodities*.

II. Results

The results of our calculation for overpayment for Iranian imports, with the lowest-cost major exporters average unit price as the base of comparison, are startling. For 1972, the overall overpayment was $539 million or 37.8 percent on the imports of the homogeneous commodities chosen. This sample of imports

represented about 60 percent of Iran's total imports and can hardly be called a small group of commodities. For 1973 and 1974, the overall results are similar but with one interesting difference. The percentage loss figures are dramatically up for 1973 to 49.4 percent and to 46.2 percent for 1974. This result is consistent with the rapid expansion of imports resulting from the oil price increases of 1973–74

The overpayment is also startling if the base is the average unit price of the low-cost exporter from whom Iran imports. In 1972 and 1973 these overpayments are somewhat higher than those obtained using the lowest-cost exporter as a base; in 1974 they are somewhat lower.

In general, one would expect that these percentage overpayment figures should be slightly lower, unless Iran already buys from the lowest-cost exporter. But such overpayments *could* be higher given that the base figures for our calculation are in fact *average unit values*; there is the possibility that exporters do not charge the same price to each importer and that Iran imports at a price below the aggregate unit value of the lowest-cost exporter in the world. A cursory glance at unit values of Indian exports (or for that matter, the results of Hufbauer and O'Neill) to different countries shows a high variance for certain categories. This being the case in many but not all instances, Iran's percentage overpayments are slightly higher.

One could object to these calculations because they may be too aggregative and thus nonhomogeneous. As a check, we calculated the 4- and 5-digit SITC imports of Iran from the *OECD* (1974). The percentage overpayment figure for Iran's imports, calculated on the basis of the low-cost producer within the *OECD* countries was 43.3 percent. Similar figures were obtained for overpayment when the base is the low-cost producer among those from whom Iran actually imported. One could still conceivably argue that such figures are for products that are not "totally" homogeneous, and thus the results may be suspect. However, the overpayment figures seem to hold up for a broad category of *very* homogeneous goods: for example, at the 3-digit level—for sugar and honey (28 percent in 1972, 54 percent in 1973, 65 percent in 1974), for tea and maté (70 percent in 1974), for maize (31 percent in 1972, 35 percent in 1973, 47 percent in 1974), for milk and cream (70 percent in 1972, 80 percent in 1973, 87 percent in 1974); for example at the 4-digit level in 1974—bovine cattle (82 percent), animal oils (48 percent), milk and cream in solid form (60 percent), glazed rice (63 percent), phosphatic fertilizers (80 percent), synthetic yarn and thread (65 percent), welded tubes and pipes of iron and steel (66 percent); and for plywood (86 percent) at the 5-digit level.

Lest Iran be considered unique among developing countries, our calculations for India (1974) at the 3-digit level also confirm these results. India's overall overpayment for her imports in the specified categories is 38.2 percent. However, India's overpayment calculated on a base of those from whom it imports is only 20.0 percent. This would indicate that India does not already buy from the low-cost producers in many categories.

Finally, we did similar calculations at the 3-digit level for Sweden (1974), a developed country. The overall overpayment figure was 47.4 percent. Again the overpayment figures decline, as in the case of India, for the calculation based on the actual set of exporting countries, to 39.0 percent.

III. Conclusion

These apparent "overpayment" figures are extremely large. The reasons outlined in the introduction could certainly account for this enormous divergence from what we would expect. As such, countries should carefully examine their patterns of trade. However, it is unlikely that any *single* factor could explain such large overpayments for imports. For instance, in the case of Iran, customs union arrangements or closed trading patterns could not explain these results. Furthermore, credit differences, though important for the exports of some countries, are unlikely to explain 40 percent deviations from low-cost producers (see *Economist*, Mar. 4, 1978, pp. 67–68). Also, given that our unit prices are FOB,

transport cost and insurance may account for *our results; again it is highly unlikely that transport costs, or even transport costs and credit terms together, could explain such large deviations in price. In fact, for Iran and India freight and insurance costs expressed as* a percentage of merchandise imports were 12 percent;[1] therefore, an average of such costs was roughly 12 percent, but the *difference* in transport and insurance costs between different import sources would be substantially less. One could also argue that the cause of this empirical curiosity may be none of our explanations but the existence of "terrible" data.

For the result regarding Iran's trade with the European Economic Community (*EEC*) another possible explanation is available. The formation of the *EEC* in 1958 led to trade diversion in agricultural goods, since the then uniformly protected members were all extramarginal producers. Let us take the case of wheat. Producers' prices for wheat in the *EEC* as a percentage of the world price reveals this inefficiency: 1958, 149 percent; 1970, 214 percent; 1971, 189 percent; 1972, 209 percent; 1973, 153 percent. Despite those extramarginal production costs, the *EEC* expanded its market share in total world wheat production. In 1968 the *EEC* produced 10 percent of the world production, 11.2 percent in 1971, and 12 percent in 1973. The degree of self-sufficiency in the *EEC* wheat production reached 116 percent in 1974–75, which makes the *EEC* a net exporter (in absence of substantial additions to storage). In 1972, the average producer price in the *EEC* was $3.10, yet the world average export unit value was $2.30. Egypt imported one million tons from the *EEC* at a price of $1.60 per bushel while Senegal was charged $1.50 per bushel. In 1974, the *EEC* exported wheat to Senegal for $2.00 per bushel, with the world market price being close to $5.00. These are potential cases of disguised foreign aid.

No matter what is behind such results, they clearly provide a reason as to why empirical tests of trade theories have been so mixed. One can also question the usefulness of empirical work on export performance which defines the residual terms as the influence of competitive forces (see J. David Richardson). In short, the validity of a great deal of empirical work on trade theory would appear to be, at best, very questionable. Finally, our results indicate that government policy may be necessary in order to achieve the lowest import bill for a given quantity of imports. Turkey has been trying to achieve such a goal by requiring importers to demonstrate that goods were purchased at the world price, but this method has resulted in an administrative nightmare.

If the results in fact reflect overpayment for imports by Iran, then this overpayment is only one dimension of the economic waste that has resulted from the oil price increases of 1973–74; with import overpayment going from 37.8 percent in 1972 to 49.4 percent in 1973. In addition industrialization has proceeded with little or no consideration given to comparative advantage and domestic factor endowments. The structure of tariffs has not been based on rational economic factors. In short, this has resulted in factories without workers, output produced in many instances at above twice the world price, and negative value-added in several industries. All of this has occurred at a rapid rate and consequently the effective rate of inflation has been in excess of 25 percent as compared to around 5 percent in earlier periods (see Askari and Cummings). In fact, an interesting comparison of the Middle Eastern development experiences can be made to that of the California gold rush era (see Askari, Cummings, and Howard Reed). In conclusion, Iran's policymakers must turn their attention to eliminating waste, and reducing overpayment for imports would be a good starting point.

[1] Derived from *Balance of Payments Yearbook*. Also these same general percentages are confirmed for certain categories of trade in *UNCTAD*, p. 60.

REFERENCES

Hossein Askari and John T. Cummings, *The Economies of the Middle East in the 1970's: A Comparative Approach,* New York 1976.

———, ———, **and H. Reed,** "Middle East:

Economic Development or Gold Rush?," paper presented at the Western Economic Conference, Hawaii 1978.

J. **Bhagwati**, "On the Under-Invoicing of Imports," *Bull. Oxford Univ. Inst. Econ. Statist.*, Nov. 1964, *27*, 389–97.

———, "Fiscal Policies, the Faking of Foreign Trade Declarations, and the Balance of Payments," *Bull. Oxford Univ. Inst Econ. Statist.*, Feb. 1967, *29*, 61–67.

G. C. **Hufbauer and J. P. O'Neill**, "Unit Values of U.S. Machinery Exports," *J. Int. Econ.*, Aug. 1972, *2*, 265–75.

Ronald W. Jones, *"Two-ness" in Trade Theory: Costs and Benefits*, Special Papers in International Economics, No. 12, Princeton 1977.

P. **Knight**, Private Communication to Hufbauer and O'Neill.

Anne O. Krueger, *Growth Distortions and Patterns of Trade among Many Countries*, Princeton Studies in International Finance, No. 40, Princeton 1977.

J. D. Richardson, "Constant-Market Shares," *J. Int. Econ.*, May 1971, *1*, 227–39.

G. C. Winston, "Overinvoicing, Underutilization, and Distorted Growth," *Pakistan Develop. Rev.*, Winter 1970, *10*, 405–21.

Economist, March 4, 1978, 67–68.

International Monetary Fund, *Balance of Payments Yearbook*, Vol. 28, Washington 1977.

UNCTAD, *World Development in Shipping, Ports and Multimodal Transport, Review of Maritime Transport, 1972*, New York, TD/B/C.4/169, Mar. 1977.

Organization for Economic Cooperation and Development, *OECD, Series C, Trade by Commodities*, Paris 1974.

United Nations, *Yearbook of International Trade Statistics*, New York, various issues.

———, *Commodity Trade Statistics*, New York, various issues.

Trade and Employment: Chile in the 1960's

By VITTORIO CORBO AND PATRICIO MELLER*

Recently policymakers in the *LDCs* have been expressing disenchantment with the employment situation that has developed in their countries, despite what are in some cases high growth rates. Part of the blame for these conditions has been placed upon internal economic policies, which it is argued, have encouraged the expansion of low-labor-intensive manufacturing activities in the modern sector and discouraged the expansion of high-labor-intensive activities in the traditional sector. But, employment in the manufacturing sector may also be affected by the type of trade strategy pursued. The two global trade strategies, import substitution and export promotion, may affect employment through one or more of the following channels: through the output mix, the factor intensity of production, and the rate of economic growth. However since we lack sufficient knowledge about the interaction between trade and growth, we are forced to consider only the employment implications of alternative trade strategies in the context of output mix and factor intensities. Furthermore, the particular manner in which the overall strategy is implemented will also affect the rate of employment. In particular, one may cite the employment effects of different effective protection rates among industrial activities, and distortions in factor prices brought about by the trade strategy.

In this paper, we examine the effect of different trade strategies on employment in the Chilean economy. First, we analyze the factor requirements of Chilean exports and import-competing manufacturing production. For this purpose we use information on factor intensities of production of the year 1967, the last year for which a census of manufacturing was made. In order to minimize the effects of trade fluctuations, we use a three-year average for trade flows, centered in 1967. This year witnessed the start of a process of export expansion which terminated in 1970. Then, also for 1967, we study how the trade strategy and its methods of implementation (especially differentiated incentives as expressed by the structure of effective protection) affected the factor requirements of that year.

Finally, to shed light on the employment effects of export expansion, we simulate the effects on factor requirements of a change in the export *mix*. For this purpose, we consider Chile's export basket of 1976–77. This period, when compared with that of 1966–68, is characterized by a major shift in trade policies from emphasis on import substitution to export promotion. We perform two simulations of the effects on factor demand of the new mix, one with the average weights of the 1976–77 period, and one with the weights of the marginal change in exports. The marginal change in exports is obtained by comparing 1976–77 annual averages with the 1966–67–68 annual averages. In both these simulations the factor intensities of individual industries are assumed equal to their 1967 level.

I. Factor Requirements of Chilean Trade in Manufactures: 1966–68

Although our main purpose is to study the trade-employment relations, as a background we review the main theoretical hypothesis regarding the pattern of trade. In a neoclassical multicountry multicommodity two-factor trade model (see Ronald Jones; Anne Krueger, 1977), the pattern of trade depends on

*Associate professor of economics and director, Institute of Applied Economic Research, Concordia University, and research associate, National Bureau of Economic Research (*NBER*); senior economist, Centro de Investigaciones Económicas para Latinoamerica (*CIEPLAN*), and research associate, *NBER*, respectively. This paper draws in part on material from our papers which were prepared as part of a project of the *NBER* on Alternative Trade Strategies and Employment Growth, carried out under the general directorship of Anne O. Krueger. We are grateful to her and Hal Lary for many discussions and suggestions on this topic and to Morton Stelcner for comments on a previous draft. We would also like to thank José M. Vrljičak for skillful research assistance.

the relative factor endowments of the trading partners. In this study, these are the *LDC*s and the most developed countries (*MDC*s).

Thus, for a country like Chile, given its relative factor endowments vis-à-vis the *MDC*s, one would expect its export basket to *MDC*s to be more labor (less capital) intensive than its production of goods which compete with imports from *MDC*s. By contrast, in its trade relations with *LDC*s (mostly *LAFTA* countries) we would expect Chile's export basket to be more capital (less labor) intensive than the basket of goods competing with imports from *LDC*s. If a third factor, "skill," is considered, then the factor intensity of a basket of tradeables will depend on the factor substitution in production among labor, capital, and skill and the relative factor endowments of the partner considered. Thus, we cannot postulate, a priori, the factor content of a basket of tradeables without further analysis of substitution possibilities.

In our analysis we use three factors: labor, capital, and skill. Labor is measured by the number of persons employed, capital is measured by the book value of fixed assets, and skill is measured by the number of "equivalent" blue collar workers less the number of persons employed. We divided eighty-two 4-digit ISIC manufacturing industries into three categories in accordance with their trade orientation, that is, exporting, import-competing, and non-import-competing industries. Industries are classified into the above three tradeable categories according to a criterion called the T_i coefficient defined as the ratio of net imports to domestic consumption of the output of the industry (see Krueger's unpublished paper). If $T_i \leq 0$ then industry i is considered an exporting one, if $0 < T_i \leq 0.75$, an import-competing one, and if $T_i > 0.75$, a non-import-competing one.

On the basis of the trade orientation coefficient, seven industries are classified as exporting, sixty-six as import-competing and nine as non-import-competing.[1] The seven exporting industries are: canning and preserving of fruits and vegetables (3113), canning, preserving, and processing of fish, crustacean, and similar foods (3114), wine industries (3132), malt liquors and malt (3133), sawmills, planing, and other wood mills (3311), manufacture of pulp, paper, and paperboard (3411), and manufacture of jewelry and related articles (3901).

For purposes of comparison, and to minimize the effect of the natural resource content, factor requirements in both exportables and import-competing products are computed per million escudos of domestic value-added (*DVA*). For import-competing products two alternative sets of weights are used—the value-added content of the import flows and the value-added content of domestic output of the industries classified as import-competing. Using the first set of weights we obtain the factor requirements of further import substitution, and with the second set of weights we obtain the factor requirements of the existing industries which produce import-competing products.

Table 1 presents the results for direct and for direct plus home-goods indirect (*HGI*) factor requirements in exportables and import-competing products. The latter measure corresponds to the so-called "Corden measure." When we analyze the results within each tradeable category, we observe substantial differences in factor requirements and factor proportions by direction of trade.

Let us first consider the results for direct requirements. For exportables, labor requirements for a basket with *MDC* weights are 2.31 (68.2/29.5) times the one for a basket with *LDC* weights. Skill requirements with *MDC* weights are only .68 times the one for a basket with *LDC* weights. The capital requirements of both baskets are fairly similar, with slightly higher requirements for the basket with *LDC* weights. If *HGI* requirements are included, the same pattern of results emerges. Labor requirements remain substantially higher for the basket with *MDC* weights (more than 2 to 1) and skill require-

[1]Copper manufacturing (ISIC 3721) is considered a natural resource based industry, and therefore is left out of our analysis. We also experimented with 0.90 as a cutoff point but the results were only marginally affected when this alternative was used.

TABLE 1—FACTOR REQUIREMENTS IN EXPORTABLES AND IMPORT-COMPETING PRODUCTS BY DESTINATION AND ORIGIN OF TRADE FLOWS: 1966–68

Weights	Labor: Number of persons employed per million escudos of *DVA*		Capital: Thousands escudos of fixed assets per million escudos of *DVA*		Skill: Number of skill units per million escudos of *DVA*		Capital-Labor Ratio		Skill-Labor Ratio	
	Direct (1)	Direct plus *HGI* (2)	Direct (3)	Direct plus *HGI* (4)	Direct (5)	Direct plus *HGI* (6)	(7) = (3):(1)	(8) = (4):(2)	(9) = (5):(1)	(10) = (6):(2)
Exports to:				Exportables						
a. World	52.8	85.6	1598.4	1717.3	115.6	77.3	30.3	20.1	2.19	0.90
b. *MDCs*	68.2	105.1	1566.0	1726.8	97.6	48.5	23.0	16.4	1.43	0.46
c. *LDCs*	29.5	50.4	1647.7	1699.9	142.9	128.8	55.9	33.7	4.84	2.55
Imports from:				Import-Competing Products						
d. World	42.6	60.1	851.5	983.4	167.6	145.6	20.0	16.4	3.93	2.42
e. *MDCs*	42.6	60.0	793.2	910.2	169.4	147.7	18.6	15.2	3.98	2.46
f. *LDCs*	42.5	60.7	1151.2	1338.5	157.3	134.3	27.1	22.0	3.70	2.21
Domestic Output:	43.4	62.1	910.8	1078.0	160.6	140.4	21.0	17.4	3.70	2.26
Exports and Imports:				Ratio of Requirements						
g. World (a : d)	1.24	1.42	1.88	1.75	0.69	0.53	1.52	1.23	0.56	0.37
h. *MDLCs* (b : e)	1.60	1.75	1.97	1.90	0.58	0.33	1.24	1.08	0.36	0.19
i. *LDCs* (c : f)	0.69	0.83	1.43	1.27	0.91	0.96	2.06	1.53	1.31	1.15

ments remain substantially lower (almost 1 to 3). For capital, by contrast, the requirements of both baskets remain fairly similar, with only slightly higher requirements for the basket with *LDC* weights.

If we consider the results by region (bottom, Table 1) we observe that in Chile's trade with both the world and with the *MDC*'s labor and capital requirements in exportables are higher than in import-competing products. By contrast, skill requirements in exportables are lower than in import-competing products. In Chile's trade with *LDC*'s, capital requirements in exportables are higher than in import-competing products, but labor and skill requirements in exportables are lower than in import-competing products. The above findings do not change when the indirect factor requirements in home-goods industries are also included.

II. Trade Distortions and Factor Requirements

In this section we analyze how policies associated with a given trade strategy have affected the factor requirements in trade observed in Table 1. Two types of distortions are considered: distortions in product markets and distortions in factor markets. Due to space limitations, detailed results are presented here only for the effect of the first type of distortion.

The structure of effective protection rates determines a system of incentives to different manufacturing activities and thus it plays a role in determining the output mix for a given basket of tradeables. In a general equilibrium model, under certain assumptions, it can be shown that the sector with the highest protection rate will attract resources. For the other sectors, nothing can be said a priori (see especially Michael Bruno). Hence, in this section, we use the structure of effective protection rates only as an indicator of the structure of incentives created by the trade regime.

During the period analyzed here, 1966–68, the import substitution process was well advanced and the trade regime discriminated in favor of import-competing activities. The average (median) rate of protection for import-competing activities was 23.0 percent

TABLE 2—FACTOR REQUIREMENTS IN EXPORTABLES AND IMPORT-COMPETING PRODUCTS BY DESTINATION AND ORIGIN OF TRADE FLOWS AND PROTECTION LEVELS: 1966–68

	Labor Number of persons employed per million escudos of *DVA*		Capital Thousands escudos of fixed assets per million escudos of *DVA*		Skill Number of skill units per million escudos of *DVA*		Capital-Labor Ratio		Skill-Labor Ratio	
Weights	Direct (1)	Direct plus *HGI* (2)	Direct (3)	Direct plus *HGI* (4)	Direct (5)	Direct plus *HGI* (6)	(7)= (3):(1)	(8)= (4):(2)	(9)= (5):(1)	(10)= (6):(2)
Exports to:				Exportables Above Median Protection Level						
World	29.7	44.4	1775.7	1852.6	154.0	151.8	59.8	41.7	5.18	3.41
MDC's	51.4	64.4	1529.8	1682.3	185.3	189.2	29.8	26.1	3.60	2.94
LDCs	28.2	43.0	1792.7	1864.8	151.9	149.1	63.6	43.4	5.39	3.47
				Exportables Below Median Protection Level						
World	65.8	103.2	1499.1	1659.3	94.0	45.3	22.8	16.1	1.43	0.44
MDCs	68.8	106.5	1567.4	1728.3	94.0	44.0	22.8	16.2	1.37	0.41
LDCs	36.7	77.2	849.4	1110.5	93.8	56.4	23.1	14.4	2.56	0.73
Imports from:				Import-Competing Products Above Median Protection Level						
World	37.5	53.8	898.4	1066.0	137.1	117.7	24.0	19.8	3.65	2.19
MDCs	36.9	53.3	782.1	931.4	136.0	116.7	21.2	17.5	3.68	2.19
LDCs	39.8	56.0	1294.1	1514.9	140.8	121.6	32.5	27.1	3.54	2.17
Domestic Output:	36.3	52.2	1031.2	1200.0	132.5	116.7	28.4	23.0	3.65	2.24
Imports from:				Import Competing Products Below Median Protection Level						
World	48.2	67.7	799.6	883.4	201.3	179.3	16.6	13.0	4.17	2.65
MDCs	48 0	66.9	803 7	888.0	201.3	180.2	16.7	13.3	4.19	2.69
LCDs	49.3	73.8	777.4	854.8	200.4	169.2	15.8	11.6	4.06	2.29
Domestic Output.	51.2	73.2	777.0	940.9	191.9	167.1	15.2	12.9	3.74	2.28

(76.3 percent), while the one for exporting activities was 2.5 percent (3.0 percent). As a result, the trade regime affected employment in the industrial sector by making import-competing activities more attractive than exporting ones. We found in the previous section that exporting activites are more labor intensive, so that we expect a negative net effect on employment from the protection system. The protection system would also affect factor requirements and factor proportions by its influence on the output mix of a basket of tradeables. Thus, if there is a positive relation between the level of protection and the capital intensity of a sector, then the protection system probably would further affect the output mix in favor of capital-intensive commodities. One cannot study fully the impact of the protection system on the output mix of each tradeable basket without developing a full general equilibrium model in which the impact of tariffs on production and demand are taken into account. This general equilibrium exercise is not attempted here. Rather, less ambitiously, within each category of tradeables, we study the association between the incentives provided by the trade regime and the factor requirements and factor proportions.

In Table 2 we have grouped the industries within each tradeable category into two sets—one with effective protection rates above and one with effective protection rates below the median of the corresponding category. Then, for each set of industries, factor requirements and factor proportions are analyzed. In analyzing these results one should bear in mind that the variance in protection rates is substantially higher for import-competing products than for exportables.

A major result is that within each tradeable category and for a given direction of trade, the bundle of goods with the protection level

exceeding the median *always* has lower labor requirements than the bundle of goods with the protection level below the median. These results lend support to the proposition that the protection system has created incentives for low labor requirements per unit of *DVA* in both tradeable categories.

In the case of capital, there is not a unique pattern. The results depend upon the direction of trade and upon whether one is considering direct or direct plus *HGI* effects. One interesting result is that when domestic output weights are used, the basket with above median protection level requires about 33 percent more capital than the one with below median protection level. From this one may conclude that the protection system created incentives to make the production of import-competing products more capital intensive than in a regime with uniform protection. For skill, the protection system creates a positive bias for exportables and a negative bias for import-competing products.

The trade regime also affects factor requirements through distortions in factor markets. During the period studied here, 1966–68, the currency was overvalued and the tariff system was such that tariffs on capital goods were lower than the tariffs on other goods. Thus, it would seem that the trade regime encouraged the use of capital-intensive techniques within each industry, and increased the profitability of capital-intensive industries above their no-distortion level. Elsewhere, we have estimated (see the authors, 1978) that employment requirements are around 6 percent lower, capital requirements around 20 percent higher, and skill requirements around 6 percent lower than in a situation without this distortion. The magnitude of these changes is certainly not negligible.

III. Factor Requirements of Export Expansion: A Simulation Experiment

The factor requirements and factor proportions analyzed in the previous two sections refer to the period 1966–68. In particular, the results for exportables refer to the existing export basket of that period. In the case of a reorientation of the trade strategy in favor of exporting activities the relevant basket of exportables to consider is the one given by the marginal exports, where due account is taken of the market for exports. To shed light on this issue we analyze here the factor requirements and factor proportions of the expansion in exports observed in the period 1976–77. This period concides with a major reduction in nominal tariffs and a substantial devaluation of the currency, and hence reflects a change in relative incentives for tradeable production in favor of exporting activities. We take the actual export levels of 1976–77 to derive the weights of a basket of exportables. In computing factor requirements of this basket we use the 1967 factor intensities per unit of *DVA*. In this way we neutralize for any effect on factor intensities due to the substantial decrease in wage-rental ratio observed when comparing the 1976–77 period with the 1966–68 period.

The results of this simulation are presented in Table 3. We observe, when comparing the results from this table with the one from Table 1, that the basket of exportables with weights equal to the *DVA* content of the marginal change in exports is more labor intensive than the one with 1966–68 export weights. The major cause of the difference is the increase in labor requirements in exports to *LDC*s. In the case of capital and skill requirements the difference is only minor.

IV. Concluding Remarks

Our analysis leads us to conclude that the trade strategy followed by Chile during the 1960's was detrimental to employment. Two general types of causes contributed to this result. First, there is the overall bias in trade incentives in favor of the low-labor-intensive import-competing manufacturing activities. Second, one may cite the distortions in product and factor markets brought about by the trade regime. These distortions affected the output mix of both exportables and import-competing activities by favoring low-labor-intensive products, and by generating a higher wage-rental ratio which in turn encouraged the use of low-labor-intensive techniques in

TABLE 3—FACTOR REQUIREMENTS IN EXPORTABLES BY DESTINATION OF TRADE FLOWS: SIMULATIONS FOR 1976–77

Weights	Labor: Number of persons employed per million escudos of *DVA*		Capital: Thousands escudos of fixed assets: per millions escudos of *DVA*		Skill: Number of skill units per million escudos of *DVA*		Capital-Labor Ratio		Skill-Labor Ratio	
	Direct (1)	Direct plus *HGI* (2)	Direct (3)	Direct plus *HGI* (4)	Direct (5)	Direct plus *HGI* (6)	Direct (7)	Direct plus *HGI* (8)	Direct (9)	Direct plus *HGI* (10)
Total Exports To										
World	58.2	87.1	1557 6	1709.0	126.3	91.6	26.8	19.6	2.17	1.05
MDCs	68.8	105.7	1554.5	1714.5	95.7	46.3	22.6	16.2	1.39	0.44
LDCs	48.6	66 9	1554.8	1685 5	153.7	139.3	32 0	25.2	3.16	2.08
Marginal Exports To[a]										
World	60.4	87.7	1540.0	1705 0	130.9	97.8	25.5	19.4	2.17	1.12
MDCs	69.3	106.0	1545.1	1704.8	94 9	45 4	22.3	16.1	1.37	0.43
LDCs	54.1	72.5	1533.2	1700.7	156 6	141 3	28.4	23 5	2.90	1 95

[a]Marginal here refers to the difference in export levels between averages for the years 1976–77 and 1966–68, both at 1967 prices

tradeable activities. The results with respect to the factor intensities of exportables are robust with respect to the weighting system used.

REFERENCES

M. Bruno, "Protection and Tariff Change Under General Equilibrium," *J. Int. Econ.*, Aug. 1973, *3*, 205–25.

V. Corbo and P. Meller, "Alternative Trade Strategies and Employment Implications: Chile" in Anne O. Krueger, Hal Lary, and Narong Chai Akrasanee, eds., *Alternative Trade Strategies and Employment Implications: Country Results*, Nat. Bur. Econ. Res., unpublished.

——— **and** ———, "Trade and Employment: Chile," unpublished paper, Nat. Bur. Econ. Res., 1978.

Ronald Jones, *Two-ness in Trade Theory: Costs and Benefits*, Special Papers in International Economics, No. 12, Princeton 1977.

Anne O. Krueger, *Growth, Distortions and Pattern of Trade Among Many Countries*, Princeton Studies in International Finance, No. 40, Princeton 1977.

———, in Anne O. Krueger, Hal Lary, and Narong Chai Akrasanee, eds., *Alternative Trade Strategies and Employment Implications: Country Results*, Nat. Bur. Econ. Res., unpublished.

The Role of a Tax-Based Incomes Policy

By LAURENCE S. SEIDMAN*

A tax-based incomes policy (*TIP*) is an innovative approach to the inflation-unemployment dilemma. First proposed by Sidney Weintraub and Henry Wallich, a *TIP* has recently begun to receive serious attention from economists (see Weintraub and Wallich, and Arthur Okun and George Perry). A *TIP* would provide a tax incentive for the employer, and/or employees, at each firm to reduce the size of the firm's wage increase. Elsewhere (1978a) I have examined the issues bearing on the optimal design of a *TIP*. In this paper, I will attempt to clarify the role of *TIP* by focusing on two important aspects: 1) the micro-economic rationale for *TIP* and 2) the compatibility of *TIP* with the monetary view of inflation.

I. The Micro-Economic Rationale for *TIP*

The classification of the inflation-unemployment problem as "macroeconomic" has long been an obstacle to analysis and policy formulation. Once a particular micro-economic view is adopted, the recommendation of a tax-based incomes policy should be a natural response for many economists.

When a problem is defined as micro-economic, economists check to discover why each economic agent is not bearing the full social cost of its actions. Once the externality is detected, a tax (or subsidy) is recommended to eliminate the divergence between private and social cost. For example, most economists regard a zero price as the cause of excess environmental pollution, and recommend an effluent tax to "internalize the externality."

*University of Pennsylvania. I am grateful to Eric Grossman who wrote the program for and executed the computer simulations.

Because the pollution problem is clearly classified as microeconomic, certain well-established principles govern the reaction of most economists. First, since an externality is present, optimality cannot be achieved without the appropriate tax (or subsidy). The market equilibrium, in the absence of proper government intervention, will not be optimal. Second, an attempt to persuade or "jawbone" economic agents to alter their behavior cannot substitute for a restructuring of financial incentives, since in the micro-economic sphere it is axiomatic that economic agents will pursue their own self-interest. Third, regulatory ceilings or "controls" are inferior to the appropriate tax because they inhibit the impact of market forces, thereby harming allocative efficiency. In contrast, the proper tax corrects the externality without altering the other market incentives operating on each agent or limiting each agent's response.

If the same micro-economic perspective is applied to the inflation-unemployment problem, a similar externality can be detected. No externality would exist in a classical, purely atomistic labor market, in the absence of trade unions, income maintenance programs, taxes, and ethical constraints on employers and workers. In such a labor market each worker would engage in a fierce wage competition with other workers. An unemployed worker would offer to work for less than the going wage if that wage exceeded the value to him of leisure or job search. Each employer would cut the wage and replace an employed worker with an unemployed worker whenever such substitution was profitable. The equilibrium unemployment rate that would emerge in such a market would be allocatively efficient.

Many would agree (including this economist) that trade unions, income maintenance

programs, government programs requiring taxes, and ethical restraints on employers and workers have on the whole significantly advanced social welfare. Nevertheless, these developments have also introduced a substantial externality into the labor market. In the modern labor market if macro policy temporarily reduces the unemployment rate to the classical equilibrium rate, these features would cause the average firm to raise its rate of wage increase. This micro-economic behavior of the average firm imposes a social cost in either of two forms. If monetary and fiscal policy attempt to maintain the classical equilibrium unemployment rate, the result is the "public bad" called accelerating inflation. If monetary and fiscal policy acquiesce in an unemployment rate sufficiently high to prevent wage inflation from accelerating, the result is above-optimal unemployment and lost output, the value of which exceeds the value of leisure or job search to the marginal unemployed.

The modern features of the labor market therefore raise the nonaccelerating inflation rate of unemployment (*NAIRU*) above what it would be in the classical labor market. With respect to allocative efficiency, a "first best" policy would be to return to the conditions of the classical labor market. Most would regard such an approach, however, as neither feasible nor desirable. Although reform of income maintenance programs and taxation may be able to reduce the *NAIRU* without harming equity, a significant externality is bound to remain as long as there are unions, ethical restraints concerning wage competition, and income maintenance programs and taxes (even if reformed).

A "second best" strategy is suggested by the recognition that these features cause the average firm to grant a larger wage increase than is optimal at any given unemployment rate. By doing so, each firm helps maintain an above-optimal *NAIRU*. It should therefore be possible to improve allocative efficiency by taxing wage increases—to internalize the external cost of raising the *NAIRU* above the classical equilibrium unemployment rate. This is the fundamental micro-economic rationale behind a tax-based incomes policy.

II. The Compatibility of *TIP* With the Monetary View of Inflation

According to the micro-economic perspective just presented, the aim of *TIP* is to reduce the *NAIRU*, not reduce inflation per se. It is therefore possible to reconcile fully *TIP* and the monetary view of inflation held by many economists. A simple macro system will now be presented in which the growth rate of the money supply governs the equilibrium inflation rate; but *TIP* governs the *NAIRU*. In this system, the proper growth rate of the money supply would be both necessary and sufficient to achieve any inflation target in the long run. Without *TIP*, however, the economy would experience a long period of "transitional" high unemployment in response to a deceleration of monetary growth. The *TIP* enables a deceleration of monetary growth to occur without causing a rise in the unemployment rate. Moreover, *TIP* should enable a permanently lower *NAIRU* to be maintained. The system is as follows:

$$(1) \quad w_t = h\left(\frac{U^* - U_t}{U_t}\right) + w_{t-1} \quad h > 0$$

$$(2) \quad p_t = w_t - a$$

where w_t = wage inflation rate; p_t = price inflation rate; U_t = unemployment rate; a = trend growth rate of labor productivity (output per man-hour); h, U^* = parameters of the economy.

Both equations can be given theoretical and econometric support (see Otto Eckstein, Michael Wachter, Perry, and the author for the wage equation; Robert J. Gordon and Weintraub for the price equation). Together they yield

$$(3) \quad p_t - p_{t-1} = h\left(\frac{U^* - U_t}{U_t}\right)$$

Since the *NAIRU* is defined as the unemployment rate at which the inflation rate (wage or price) remains constant, from (1) or (3) it is clear that the *NAIRU* is initially $U_t = U^*$.

The aim of a permanent *TIP* is to induce each firm, through a tax incentive, to grant a smaller wage increase *each period* than it otherwise would have, given the unemploy-

ment rate, and the recent wage inflation rate. A micro-economic analysis of *TIP*'s impact is given elsewhere (see the author, 1978a). If *TIP* succeeds, it will shift down the wage equation:

$$(1\text{-}TIP) \qquad w_t = h\left(\frac{U^* - U_t}{U_t}\right) + w_{t-1} - T \quad T > 0$$

where T = shift due to *TIP*.

For example, if *TIP* succeeds in causing the average firm to grant 2 percent less than it otherwise would have, then T would equal 2 percent and (3) would become

$$(3\text{-}TIP) \quad p_t - p_{t-1} = h\left(\frac{U^* - U_t}{U_t}\right) - T$$

From (1-*TIP*) and (3-*TIP*), it is clear that $U_t = U^*$ would now cause wage and price inflation to decline. The new *NAIRU* under *TIP* can be obtained by setting $(p_t - p_{t-1})$ equal to zero, and solving for U_t:

$$(4) \qquad U_t = \left(\frac{h}{h + T}\right) U^* < U^*$$

If a permanent *TIP* succeeds in shifting down the wage equation permanently, it would reduce the *NAIRU*. The above wage-price system, however, can only determine the relationship between inflation and unemployment, and *TIP*'s impact on this relationship. It cannot in itself determine either the inflation rate or the unemployment rate, because (3) or (3-*TIP*) is one equation with two unknowns, p_t and U_t. To complete the system, the growth rate of nominal aggregate demand must be specified, thereby introducing a relationship between the growth rate of real output and the inflation rate. Since the growth rate of real output can be linked to the unemployment rate through the aggregate production function, this provides the second relationship between p_t and U_t that is required to complete the system.

A monetarist assumption concerning the growth rate of nominal aggregate demand will be accepted here to emphasize the compatibility of *TIP* with the monetary view of inflation. It will be assumed that the growth rate of the income velocity of money in (5) is constant, so that the growth rate of the money supply determines the growth rate of nominal income:

$$(5) \qquad m_t + v_t \equiv p_t + q_t$$

where m_t = growth rate of the money supply, v_t = growth rate of the income velocity of money; p_t = inflation rate; q_t = growth rate of real output.

Given v_t, m_t fixes the sum of p_t and q_t. The unemployment rate, U_t, can be linked to q_t as follows:

$$(6) \qquad n_t \equiv q_t - a_t^*$$

where n_t = growth rate of employment; a_t^* = growth rate of output per worker (a_t is the growth rate of output per man-hour).

$$(7) \qquad e_t \equiv n_t - l_t$$

where e_t = growth rate of the employment rate; l_t = growth rate of the labor force.

$$(8) \qquad e_t \equiv \frac{E_t - E_{t-1}}{E_{t-1}} \equiv \frac{U_{t-1} - U_t}{1 - U_{t-1}}$$

where $E_t \equiv 1 - U_t$.

Define x_t as follows:

$$(9) \qquad x_t \equiv m_t - (a_t^* + l_t - v_t)$$

Combining (5)–(9) yields:

$$(10) \quad U_t - U_{t-1} = (1 - U_{t-1})(p_t - x_t)$$

If a_t^*, l_t, and v_t are given, x_t is determined by m_t. Given m_t and x_t, equations (3) and (10) constitute two equations with two unknowns, U_t and p_t, and thus determine U_t and p_t.

Suppose m_t is held constant at m by the monetary authorities, and that a_t^*, l_t, and v_t are all constant at a^*, l, and v, so that $x_t = x$. Then $p_t = x$ would hold the unemployment rate constant in (10); and $U_t = U^*$ would hold the inflation rate constant in (3) (prior to *TIP*). Thus, these are equilibrium values.

Will the system converge to these values under a constant m if p_o and U_o are initially not at x and U^*? If (10) is solved for p_t, and then for p_{t-1} (by lagging each U term and x one period), and these values are substituted into (3), we obtain

$$\text{(11)}\quad \frac{U_t - U_{t-1}}{1 - U_{t-1}} - \frac{U}{1 - U_{t-2}} = h\left(\frac{U^* - U_t}{U_t}\right)$$

a second-order, non-linear difference equation. I have been unable to provide, or find, a general proof that, for any initial values, U_o and U_1, the system converges to its equilibrium value, U^* (and therefore also p_t to x). Nevertheless, I ran computer simulations for (3) and (10) for alternative values of p_o and U_o. In every case, each variable oscillated, eventually converging to its equilibrium value.

In this macro system, therefore, it appears that for any initial values U_o and p_o, if money growth is held constant at m, then the inflation rate will eventually stabilize at x, given by (9); and the unemployment rate, at U^* (prior to *TIP*).

If *TIP* is permanently introduced, (3) becomes (3-*TIP*), and the equilibrium unemployment rate—the *NAIRU*—will be reduced according to (4), yielding

$$\text{(11-}TIP\text{)}\quad \frac{U_t - U_{t-1}}{1 - U_{t-1}} - \frac{U_{t-1} - U_{t-2}}{1 - U_{t-2}} = (h + T)\left(\frac{NAIRU - U_t}{U_t}\right)$$

where $NAIRU \equiv [h/h + T]U^*$ (given by (4)).

The *TIP* should not, however, alter the derivation of (10). Thus, the equilibrium inflation rate would remain x, which is determined by m. In this macro system, therefore, the role of *TIP* is to lower the *NAIRU*. The relationship between the growth rate of the money supply and the equilibrium inflation rate would be unaffected by *TIP*.

The significance of introducing *TIP* in this macro system can be seen in the numerical example presented in Table 1. Suppose initially $U_o = U^* = 6$ percent, and $p_o = 6$ percent. Assuming $a = 2$ percent, $l = 1$ percent, $v = 2$ percent, an m of 7 percent would maintain this equilibrium, since x would be 6 percent, so that $U_1 = 6$ percent and $p_1 = 6$ percent would satisfy both (3) and (10). In the absence of *TIP*, for $h = .04$, the

TABLE 1—MONETARY DECELERATION WITH AND WITHOUT *TIP*

		With *TIP*		Without *TIP*	
Period	m_t	U_t	p_t	U_t	p_t
0	7.0	6.0	6.0	6.0	6.0
1	5.7	5.6	4.3	6.8	5.5
2	4.3	5.2	2.9	8.0	4.6
3	3.3	4.8	1.9	8.8	3.3
4	2.3	4.6	1.1	9.3	1.8
5	1.8	4.4	0.6	9.1	0.5
6	1.5	4.2	0.3	8.1	−0.5
7	1.2	4.1	0.1	6.9	−1.1
8	1.0	4.1	0.0	5.9	−1.0
9	1.1	4.0	−0.1	5.4	−0.6
10	1.0	4.0	0.0	5.3	−0.1
11	1.0	4.0	0.0	5.5	0.2
12	1.0	4.0	0.0	5.9	0.3
13	1.0	4.0	0.0	6.1	0.3
14	1.0	4.0	0.0	6.2	0.1
15	1.0	4.0	0.0	6.2	0.0

Note: Derived from the model under the following assumptions: $U = 6.0$ percent, $h = .04$, $T = 02$; $a^* = 2$ percent, $l = 1$ percent, $v = 2$ percent, so $x_t = m_t - 1\ 0$ percent

gradual deceleration of m_t from 7 to 1 percent (x_t from 6 to 0 percent) would produce the paths shown in the table (calculated by solving (3) and (10) period by period). After a significant rise in U_t above 6 percent, oscillations would eventually bring convergence to the 6 percent *NAIRU*, and a 0 percent inflation rate.

If *TIP* is introduced with a T of 2 percent, then the identical deceleration of m_t (and x_t) would produce the paths shown (calculated by solving (3-*TIP*) and (10) period by period). In the presence of gradual monetary decleration, the role of *TIP* is: (a) to avoid a "transitional" rise in the unemployment rate and (b) to cause the macro system to stabilize permanently at a lower unemployment rate.

Because the growth rate of the money supply controls the equilibrium inflation rate in this system, it would be more precise to describe the role of *TIP* to be the reduction in the nonaccelerating-inflation rate of unemployment, not the reduction in inflation per se. Since the transitional rise in the unemployment rate without *TIP* may well be judged to be socially unacceptable, and the final

NAIRU without *TIP* may be viewed as a reward unworthy of the transitional sacrifice, monetary deceleration may not be undertaken without *TIP*. In this important practical sense *TIP* deserves to be regarded as an anti-inflation policy: it makes a policy of monetary deceleration feasible and desirable. If *TIP* does succeed in reducing the *NAIRU*, it also should be regarded as an anti-unemployment policy.

It is important to understand how this analysis of the role of *TIP* relates to the wage view of inflation with which *TIP* is usually associated. In our system, it is also true from (2) that the price inflation rate is a function of the wage inflation rate, through the well-established price-unit labor cost linkage. The wage inflation rate, however, is endogenous in our system, given by (1), while the money supply growth rate is capable of being exogenous. *TIP* is of course also consistent with an exogenous wage theory of inflation (see Weintraub). Our system shows, however, that *TIP* and the wage and monetary views of inflation can all be reconciled.

REFERENCES

Otto Eckstein *The Econometrics of Price Determination*, Washington 1972.

R. J. Gordon, "The Impact of Aggregate Demand on Prices," *Brookings Papers*, Washington 1975, *3*, 613–62.

A. P. Lerner, "Stagflation—Its Cause and Cure," *Challenge*, Sept./Oct. 1977, *20*, 14–19.

A. M. Okun, "The Great Stagflation Swamp," *Challenge*, Nov./Dec. 1977, *20*, 6–13.

Arthur M. Okun and George L. Perry, *Brookings Papers on Economic Activity 2*, Special Issue: *Innovative Policies to Slow Inflation*, Washington 1978.

G. L. Perry, "Determinants of Wage Inflation Around the World," *Brookings Papers*, Washington 1975, *2*, 403–35.

L. S. Seidman, (1976a) "A New Approach to the Control of Inflation," *Challenge*, July/Aug. 1976, *19*, 39–43.

———, (1976b) "A Payroll Tax Credit to Restrain Inflation," *Nat. Tax J.*, Dec. 1976, *29*, 398–412.

———, (1978a) "Tax-Based Incomes Policies," *Brookings Papers*, Washington 1978, *2*, 301–48.

———, (1978b) "Would Tax Shifting Undermine the Tax-Based Incomes Policy?," *J. Econ. Issues*, Sept. 1978, *12*, 647–76.

———, "The Return of the Profit Rate to the Wage Equation," *Rev. Econ. Statist.*, Feb. 1979.

———, "A Note on the Hazards of the Monetarist Rule," *Southern Econ. J.*, Jan. 1979.

M. L. Wachter, "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers*, Washington 1976, *1*, 115–59.

H. C. Wallich and S. Weintraub, "A Tax-Based Incomes Policy," *J. Econ. Issues*, June 1971, *5*, 1–19.

Sidney Weintraub, *Capitalism's Inflation and Unemployment Crisis*, Reading 1978.

Comparing *TIP* to Wage Subsidies

By DONALD A. NICHOLS*

This paper derives some analytic results concerning the possible effects of a tax-based incomes policy (*TIP*), and compares them to the effects of a wage subsidy or a decreased payroll tax. The policies are compared using a model of firm equilibrium which is somewhat simpler than that of Yehuda Kotowitz and Richard Portes, and R. W. Latham and David Peel. Because the model is one of firm equilibrium, it ignores both interactions among firms and workers, and the bargaining process. As a result, it cannot answer all possible questions about the effectiveness of a *TIP*. Nevertheless, the model can address an important question that lies at the very heart of the issue of the possible effectiveness of a *TIP*: in what way would a *TIP* influence a firm to change its wage and price decisions, assuming nothing else in the economy were to be changed.

If, as shown below, *certain versions of TIP appear no more effective than wage subsidies in changing firm behavior,* then it can well be questioned whether the addition of union behavior and firm interactions would change this result. Furthermore, it is possible that a *simple reduction in payroll taxes* could accomplish the same results as a *TIP*. (There are, of course, many possible versions of *TIP*, and not all of them can be replicated exactly by payroll taxes or subsidies.)

To show that a *TIP* may be identical to other employment taxes or subsidies is of course not a criticism of either policy. Rather, it indicates the need for coordinating policies that have an effect on the demand for labor in order to avoid their working at cross purposes, such as in the example given below of a *TIP* and a wage subsidy that completely offset each other.

*University of Wisconsin; U.S. Department of Labor.

I. A Profits Tax *TIP*

The most widely reported versions of the *TIP* are those in which the corporate tax rate is made to vary with the rate of wage increase offered by the firm. These include Arthur Okun's (1977) plan, which would lower the corporate tax rate for firms that offer wage increases below some standard, and Henry Wallich and Sidney Weintraub's plan, which would raise tax rates for noncomplying firms. Laurence Seidman (1976) reports a symmetric plan that would incorporate both a "carrot" and a "stick."

The following notation and functional relations are used: revenues R are assumed to be a function of employment L; the profits tax rate t is a function of the wage rate w; pretax profits equal revenues minus labor costs. These features are combined to define after-tax profits (π):

$$(1) \qquad \pi = [1 - t(w)][R(L) - wL]$$

To make the wage decision interesting, it is assumed the firm faces an upward-sloping labor supply function:

$$(2) \qquad L = L(w); \quad L' > 0$$

The firm's profit-maximizing wage and employment conditions are derived from the Lagrangian:

$$(3) \qquad \pi = [1 - t(w)][R(L) - wL] - \lambda[L - L(w)]$$

$$\partial\pi/\partial L = [1 - t(w)][R' - w] - \lambda = 0$$

$$\partial\pi/\partial w = -t'[R(L) - wL] - [1 - t(w)]L + \lambda L' = 0$$

These conditions can be easily compared to the conventional results by solving them for R'. This is done in (4) where the elasticity of the supply of labor with respect to the wage rate is represented by the letter e:

$$(4)\quad R' = w[1 + \frac{1}{e}] + \frac{t'}{1 - t}[\frac{R(L)}{L} - w]\frac{w}{e}$$

Note that in the absence of a *TIP*, i.e., $t' = 0$, equation (4) yields the usual relation between the marginal revenue product of labor R' and the wage rate w: in the case where $e = \infty$, R' is to be set equal to the wage; otherwise it also depends on the elasticity of the labor supply function.

There is some question of how to judge the effectiveness of *TIP* in this model. Since the labor supply function slopes upward, the only way to hold down wages is to restrict output. But this would cause prices to rise, the exact opposite of what is intended by the program. This awkward feature of the model was emphasized by Latham and Peel, who argue that a *TIP* could actually be inflationary. Whether this possibility is best attributed to the simple-minded nature of the model or whether it points out a fundamental flaw in *TIP* is a matter of judgment. I discuss this issue below. Since I am interested in issues affecting the demand for labor, I will assume the *TIP* is effective if it reduces the wage offered by the firm. This effect can be seen in equation (4).

First, as might be expected, a *TIP* will have no influence on a firm that faces a perfectly competitive labor market ($e = \infty$). This result is due, of course, to the partial equilibrium nature of the model. When aggregated over competitive firms, it might well be that the *TIP* could influence the overall wage rate.

Second, the effectiveness of *TIP* will be greater if the *TIP* is marginal in nature. This is how I interpret the role of the term $t'/(1 - t)$. It would be hard to have t' be large if t is a linear function of w. What is needed is a *TIP* that changes tax rates by a large amount in the small neighborhood of w that is relevant for purposes of inflation control. The Okun and the Wallich-Weintraub *TIPs* incorporate this characteristic with a vengeance. They are discontinuous at some threshold rate of wage increase. The Seidman version on the other hand is effectively linear.

Third, the effectiveness of a *TIP* will be related to the firm's profit per worker, $(R(L)/L - w)$. (See Slitor 1978.) This is due to the fact that the size of the incentive offered to the firm through a cut in its profit tax rate varies with the amount of profit the firm is making. Thus unusually profitable firms or capital intensive firms would have the greatest incentive to restrict wages under a profits tax based *TIP*.

Only the third characteristic would be a major drawback for a profits based *TIP*. It probably would be considered unfair to have the *TIP* impose greater limitations on the wages of profitable, capital-intensive firms than on those of other firms. Of course it would be necessary to examine the distribution of firms by their profit per worker to see just how serious this drawback might be operationally. Moreover Okun (1978) has suggested that the problem could be removed by scaling the *TIP* computation with some measure of labor intensiveness, preferably taken from the previous year's financial data. Another possibility, pointed out by Slitor, would be to base the *TIP* on labor compensation rather than on profits.

II. A Compensation Based *TIP*

A compensation based *TIP* would rest evenly on each dollar of labor compensation, meaning that a firm would have twice the incentive to hold down the wages on a high wage worker as on one earning half as much. A profits based *TIP* would have this same incentive, of course, if the wage rate used for tax purposes was simply an unweighted average of the wages paid to all classes of employees.

Rather than apply the *TIP* to corporate profits, a compensation based *TIP* would be defined as a tax on some percentage of the wage increase in excess of the chosen standard. A special case of this general form is the proposal of Slitor, who would deny firms the right to deduct the excess wage payments as a cost when computing income tax liability. In this case, the penalty rate on wage payments

would be the corporate tax rate itself. This case is shown in equation (5) where w_o is the wage that would result from following the wage guideline.

$$\text{(5)} \qquad TIP\ Tax = t(w - w_o)L$$

Looking at equation (5), it becomes clear that this particular compensation based *TIP* looks a great deal like a payroll tax or a negative wage subsidy or employment subsidy. To see how much these policies have in common, it is useful to include them all in the model being developed.

Let s denote a general wage subsidy expressed as a percentage of wage payments. Wage subsidies are often marginal in nature, but the marginality usually refers to changes in employment rather than in wage levels. That is, in the common form of wage subsidy firms are eligible for the subsidy only for increases in employment beyond some base, L_o. For this reason they are often called employment subsidies, for example the existing employment tax credit.

$$\text{(6)} \qquad Wage\ subsidy = sw(L - L_o)$$

According to (6) the firm would have to pay a penalty for reductions in employment below the base. This characteristic is not a part of any of the proposed wage subsidies, but it is maintained for symmetry with the treatment of *TIP*. The effect of taxes that are not symmetric can be seen from the results about to be derived simply by ignoring the effects that take place above or below the employment or wage base.

The maximization problem for a firm facing the *TIP* of equation (5) and the wage subsidy of equation (6) is expressed in (7), and the solution to the problem is shown in (8).

$$\text{(7)} \quad \text{Max } \pi = R(L) - wL + sw(L - L_o) - t(w - w_o)L - \lambda[L - L(w)]$$

$$\partial\pi/\partial L = R' - w(1 + t - s) + tw_o - \lambda = 0$$

$$\partial\pi/\partial w = -L(1 + t - s) - sL_o + L'\lambda = 0$$

$$\text{(8)} \quad R' = w(1 + \frac{1}{e})(1 + t - s) + \frac{swL_o}{eL} - tw$$

Comparing equation (8) to (4) we see:

1) *A compensation based TIP would be effective for firms facing competitive labor markets while a profits based TIP would not.* That is, the tax terms do not disappear from (8) when $e = \infty$, as they did from equation (4).

2) *A symmetric, compensation based TIP would tend to lower wage rates that would otherwise be above $w_oe/(e + 1)$ and to raise those that would be below.* For competitive firms, only wages above w_o would be lowered. This means the choice of a guideline w_o is important even for a symmetric *TIP*. Visually, this means the effect of the *TIP* is to rotate the demand curve for labor around the w_o point.

3) *A wage subsidy and a compensation based TIP work in opposite directions on the demand for labor.* This result, of course, depends on the choice of w_o, L_o, s, and t. But in the special case where $e = 1$ and where L_o and w_o are chosen to satisfy $L_o = L(w_o)$, the effects are exactly opposite. In that case, if s and t were to have the same absolute value, they would offset each other completely.

It should be noted that the effect of a change in the payroll tax can also be inferred from (8) by looking at the *TIP* tax for the special case where $w_o = 0$. Similarly the effects of an average wage subsidy can be inferred from the results for marginal subsidies by looking at the case where $L_o = 0$. An average wage subsidy and a payroll tax have, of course, the opposite effect on the demand for labor.

III. Implications and Some Remaining Issues

It is important to interpret correctly the result concerning the similarities between a *TIP* and wage subsidies. The important result for policy purposes is not that under certain highly restrictive conditions one version of *TIP* and one version of a wage subsidy are exactly the same or exactly the opposite.

Rather the important point to note is that both policies work by shifting the demand for labor and therefore that the two policies should be coordinated. Furthermore, where administrative problems may argue against the adoption of any particular policy, it may be possible to reach the same goals by some other employment policy that is easier to administer. On the other hand, where certain versions of the *TIP* have been proposed with key threshold values—for example, at a 6 percent wage increase where a major change in tax rates would take place—it would be difficult to pinpoint that threshold with a wage subsidy. Similarly, limitations on allowable wage payments for wage subsidies, payroll taxes, etc. might make such policies effective only for low wage workers. Since the *TIP* also would encourage employment of low wage workers, it is possible to imagine the two policies offsetting each other on wage levels but reinforcing each other in their effect on employment.

A major qualification of the results is that the model does not permit prices and wages to be influenced in the same direction. The revenue function is consistent with downward sloping or flat demand curves, while the labor supply curve slopes upward. This means that a restriction in output would always be associated with a reduction in wages, but not in prices, while an increase in output would always be associated with the reverse. This peculiarity makes it difficult to conclude which actions are truly anti-inflationary. Is it anti-inflationary to hold down wages or prices?

The answer depends, of course, on the influence that the decisions of the firm would have on other wage and price decisions, particularly in the future. If, for example, the supply of labor in this model were written as a function of wages and prices elsewhere, then it would be necessary to ask whether the changes in those wages and prices caused by any particular policy would cause the supply of labor to rise or fall. This is an empirical question about which there is no uniform agreement.

Because wages and prices cannot move in the same direction in this model, a *TIP* penalty on labor compensation—which reduces wage rates—will cause prices to rise if demand varies with price. Similarly, a *TIP* subsidy will cause prices to fall but wages to rise. Depending on one's view of the inflationary process, one or the other of these proposals could be inflationary.

It should be noted that the profits tax *TIP* with a threshold would not lead to wage increases. Below the threshold, there is no reason to raise wages, while above the threshold there is no incentive to reduce them.

The model ignores many issues of the bargaining process including those considered by Seidman (1978) and Wallich and Weintraub. The bargaining process is not easily modeled. However, one can argue that certain qualitative effects are more or less plausible with a *TIP*.

Viewed from the union's vantage, it is plausible that anything that increases profits would increase wage demands. This means that *TIP* subsidies—even those with threshold levels—would encourage higher wage demands over some part of the range of possible wage increases. The union would have no incentive to restrict its own demands, knowing that the firm would receive a tax bonus on top of the increase in profit that would normally result from paying lower wages.

It is also plausible that the source of the increased profits—whether from a *TIP*, a wage subsidy, or a payroll tax cut, whether from a marginal, an average or a discontinuous program—would have little or no effect on the union's demands. Since unions can make all-or-nothing bargains, even lump sum transfers received by the firm could influence their behavior. From the union side, a *TIP* that increases taxes paid by firms is likely to be more effective in reducing wage offers than one that reduces taxes.

Some *TIP*s specify that payments would be made through the income tax to workers in complying firms. This would shift the labor supply curve and permit both wages and prices to be moved in the same direction. It is important to distinguish this form of payment from an insurance tax cut that would be made if the target rate of inflation under the *TIP* were not met. The insurance cuts are designed

to guarantee the real purchasing power of workers who comply with the *TIP*. These insurance tax cuts, however, could be designed to accompany any anti-inflation program, and should not be considered exclusively as part of a policy that taxes or subsidizes firms based on the wages they pay.

In summary, *TIP* is one of a variety of policies that can influence the demand for labor. If we are to be sure that these policies do not work at cross purposes, they should be analyzed simultaneously. A simple example of such an analysis was developed above in which a *TIP* was shown to have an effect similar to that of a marginal payroll tax and opposite to that of a wage subsidy. It is not clear that these results would entirely disappear in more complex models.

REFERENCES

Y. Kotowitz and R. Portes, "The 'Tax on Wage Increases,' " *J. Publ. Econ.*, May 1974, *3*, 113–32.

R. W. Latham and D. A. Peel, "The 'Tax on Wage Increases' When the Firm is a Monopsonist," *J. Publ. Econ.*, Oct. 1977, *8*, 247–53.

A. M. Okun, "The Great Stagflation Swamp," *Challenge*, Nov./Dec. 1977, *20*, 6–13.

———, "Comment" on Seidman, *Brookings Papers*, Washington 1978, *2*, 353–57.

L. S. Seidman, "A New Approach to the Control of Inflation," *Challenge*, July/Aug. 1976, *19*, 39–43.

———, "Tax-Based Incomes Policies," *Brookings Papers*, Washington 1978, *2*, 301–48.

R. E. Slitor, "Tax-Based Incomes Policy; Technical and Administration Aspects," report prepared for the Board of Governors, Fed. Reserve System, Mar. 20, 1978.

H. C. Wallich and S. Weintraub, "A Tax-Based Incomes Policy," *J. Econ. Issues*, June 1971, *5*, 1–19.

Implementation and Design of Tax-Based Incomes Policies

By RICHARD E. SLITOR*

The pioneer proposal by Henry Wallich and Sidney Weintraub marks an important step in the evolution of a more effective high employment policy: the application of microeconomic incentives that discourage inflationary wage decisions and in effect internalize their social costs in order to achieve macroeconomic goals. Evolving from this innovative tax-based incomes policy (*TIP*) concept we have a wide variety of plans and variants, some of which have been published in the special Brookings volume, including Arthur Okun's tax bonus *TIP*; Laurence Seidman's hybrid approach to reduce the nonaccelerating-inflation rate of unemployment (*NAIRU*); Abba Lerner's wage increase permit plan (*WIPP*); and now the all-purpose plan embodied in the legislative draft developed by the Senate Committee on Banking, Housing, and Urban Affairs. Conceptually productive, the new pluralism complicates a compact analysis of design issues. It would be impossible for me to cover this waterfront. This discussion highlights major areas of design and implementation with emphasis on the Wallich-Weintraub wages *TIP*.

I. The Carrot vs. Stick Approach

First, a few words on the now familiar problem of choice between the penalty or stick approach exemplified by the original Wallich-Weintraub plan—still the major contender in the field—and the carrot approach, of which the Okun tax bonus plan is the prototype. The bonus appeals to considerations of political strategy and coats needed fiber with sugar that in an inflationary economy would probably have to be forthcoming anyway in the guise of a nominal tax cut. Actually, both carrot and stick involve the penalty and reward, explicit or implicit, of a dual tax level—one for guideline conformers, the other for nonconformers.

A critical consideration in choosing between carrot and stick is one of compliance and administration. The stick *TIP* can be limited to the large corporation, strategic wage settlement centrum from which the wage-cost-price spiral emanates. By contrast, the carrot system almost inevitably calls for burdensome universal coverage. Where tax bonuses are being distributed, possibly for only a decent modicum of restraint, it would be unfair to exclude large sectors of private enterprise on the grounds that they were not large enough to count or sufficiently resourceful to handle the qualifying procedures. The reward method for employers or employees, or both, also involves the technical problem of implementing it currently on a before-the-fact basis via withholding or estimated current payment and then making subsequent corrections.

II. The Coverage Issue

The allegation of unacceptably high costs involved in applying *TIP* to millions of unincorporated enterprises and small corporations is a bogy unless the plan embodies the tax bonus approach. The penalty method limited to the 2- or 3,000 largest corporations in terms of asset size would embrace about half of total corporate receipts and roughly two-thirds of total corporate assets and net earnings.

The administrative cost of a reasonable *TIP* penalty tax plan for the large corporate sector would be moderate. A figure of several million dollars is sometimes suggested by *TIP* proponents, and I do not regard it as unreasonable. Still, the uneasy skeptic may wonder what an outside figure (on the high side) would be. For a guesstimate of costs representing an outside figure, assume special

*Economic consultant.

audit of compensation increases for some 2,000 large corporate entities, with less intensive attention to perhaps 1,000 more in the next size layer down. Even if the average cost ran as high as $7- or $8,000 annually per entity, the total outside figure only would amount to $20 or $25 million. Costs would be less after the start-up period and in a less inflationary economy with more firms clearly within the guideline limit.

III. Timing Questions

The *TIP* design presents a number of questions of timing and duration. Should *TIP* be regarded as a temporary measure or as a more or less permanent addition to the fiscal and monetary policy armory?

Wage pressures have been aggravated by the jockeying for relative economic positions in the turmoil of persisting inflation catch ups and inflationary expectations. But they are inherent in the system of productivity-increase sharing via money wages, the pervasive drive for continuing improvement in living standards, and the cushioning of the human hardships of unemployment. The *TIP* should therefore be regarded as a permanent and synergistic addition to macro-economic management tools, not merely as a one-shot stabilizer designed to break the vicious circle and bring the system back to its senses.

The application of the *TIP* penalty for only a year or two would have arbitrary and inequitable impacts, depending upon particular firms' positions in the continuing round of wage adjustments to reflect productivity and cost-of-living catch ups. These differentials and associated pressure for relief would tend to be washed out under a continuing *TIP* program. Special safeguards might be needed against multiyear contracts featuring deferred wage increases which would merely build up a postponed inflation that would later inundate the system.

Another question in the timing area: Should a *TIP* penalty once incurred apply for one year only; for a fixed term of, say, three to five years; or on a cumulative basis for a lengthy or indefinite period? The options embody in varying degree the principles of continuous economic accountability vs. let-bygones-be-bygones or wipe-the-slate-clean.

Payment of excess wages may be thought of as a one-time infringement of stabilization guidelines and sufficiently penalized by a one-year sanction. Taxation beyond that time frame, cumulating one-year's penalty upon another's may be regarded as unnecessarily punitive for an economic *fait accompli*. Moreover, a cumulative penalty may destroy old businesses in an unfavorable *TIP* posture and favor new businesses—pseudo or real—that could escape such hangovers. The mechanics of a cumulative approach become quite formidable. On the other hand, there is merit in the concept of applying sustained cumulative pressure to keep wages within the guideline. Quick release from the penalty for a particular year's infringement may invite annual flirting with the guideline's limitations.

IV. Definition of the Accounting Unit

Another standard question is whether the basic accounting unit for computing excess compensation should be the corporation, the affiliated corporate group or conglomerate, the division, the plant or establishment, the wage bargaining unit, if any, or some other classification of labor.

The income tax paying unit would almost necessarily be the unit for purposes of assessing and collecting an income tax oriented *TIP* tax penalty (rate or disallowance of deduction), but another concept—the *TIP* accounting unit—is relevant to the design and implementation of this form of tax.

Asking this and related questions about the extent of aggregation or disaggregation in measuring wage increase experience does not imply serious practical problems but suggests significant design options. Disaggregation below the level of the firm or affiliated group might be theoretically acceptable under an excise-type tax penalty on excess compensation. But business firms would regard even this as unfair if their overall wage experience was within the guideline. Disaggregation in the first instance may be reversed if the results of sectoral excess wage computations,

plus and minus, are recombined for the larger income tax unit. The results of reaggregation are similar and in some cases identical with aggregate average pay increase calculations. Disaggregation by wage bargaining units is most consistent with the view of *TIP* as monitoring a social compact among labor, employer, and government, but I doubt its consistency with the general spirit of the *TIP* incentive.

Some analysts are troubled by the incongruity of combining unrelated types of business activity under a rationale of *TIP* as an incentive entering tangibly and measurably into the wage bargaining process. This concept may seem to suffer if a particular wage bargain affects *TIP* liability only through its interaction with other wage determinations throughout a heterogeneous industrial empire. However, it does not seem too much to expect modern business management to be aware of firm-wide wage developments, to respond coherently, and to apply the discipline of noninflationary wage policy without undue regard to industrial heterogeneity.

V. Form of the Tax Penalty

Use of either the income tax rate or denial of deductibility ties the tax penalty to the income tax status of the firm. The rate method ties the penalty to the amount of net earnings as well as to the fact and proportionate amount of excess compensation. A link to net earnings naturally reduces the predictability of the relationship between the amount of the excess compensation and the penalty tax, and generates differential treatments, problems which would be corrected by the denial of deduction method.

An incentive tax feature based on net income tends to discriminate against equity capital-intensive operations, with distorting effects including encouragement of debt financing. Denial of deductibility tends to avoid this bias, but has been criticized for placing the entire penalty onto the marginal cost of production. The marginal cost criticism is of considerable moment if one attaches weight to the interplay of the Vinerian cost curve system and the revenue curves in the price and production decisions of large corporations as affected by labor costs including *TIP* arising from inflationary wage settlements. However, the basic announcement effect of the *TIP* penalty is linked to the threat of triggering the *TIP* penalty, and this threat would tend to have broadly similar wage bargain effects whether via tax rate or denial of deduction. Nor is the attribution to variable costs of the *TIP* penalty as a possibility or a realized fact all bad; it contributes to the short-run substitution effects of a *TIP* (more abundant, lower-cost labor for scarcer, higher-cost labor, and some forms of mechanization) which seem to be part of its stabilizing role. I concede that this whole matter of *TIP* design effects involving cost calculations, pass through into prices, resource substitution, and impacts on employment and investment calls for further study.

Another design issue is the choice between a continuous system of gearing *TIP* penalties to all wage increments and a hurdle or threshold format in which the *TIP* reward or penalty depends, possibly on an all-or-nothing basis, upon guideline conformance. A fully continuous method applies some wage restraint incentive to all covered firms and the whole range of wage determinations. This increases the efficiency of the program but requires exact measurement of wage increases for all covered firms. The hurdle or threshold approach in its extreme form would restrict the more exacting measurement and enforcement tasks to those firms that are near the prescribed threshold. A modified threshold method which has suitable notch and scaling provisions adjusting the penalty to the degree of excess above the guideline is a preferred compromise.

An excise type of tax on excess wage increases would divorce the *TIP* penalty from income tax complications but at present does not seem to be in the mainstream of *TIP* development.

VI. Measurement of Employee Compensation and Excess Increases

Problems of wage measurement, labor classification, labor unit definition, weighting,

and pay average or index number construction have been discussed by Larry Dildine and Emil Sunley in their recent Brookings panel paper and in my study for the Federal Reserve. It is not feasible to develop this area in any detail here.

Fringe benefit identification and measurement, as part of the costing-out of wage settlements, is not a novel operation but part of rational evaluation of wage settlements. The *TIP* would call for greater alertness to novel ploys and gambits and standardized methodology in this field. In particular, the pension area constitutes a large potential for nonwage compensation and presents intellectual and practical problems. Still it seems doubtful *TIP* avoidance effort adds much to the existing incentives for employers to pay compensation of a nature not currently recognized as taxable to the employee beneficiary or under payroll taxes.

Avoidance and evasion devices under *TIP* applied via a weighted-average pay increase method would include efforts to disguise pay increases as promotions—a familiar phenomenon in organizations confronted with pay freezes. The effect of paper promotions may be further enhanced if the promoted individual's pay, despite an absolute increase, has the effect of lowering the average pay in his new class. However, in view of the gross bias of the unweighted average approach, a weighting approach would be clearly preferable.

Even rather firm opponents of *TIP* on both conceptual and workability grounds (for example, Alan Greenspan) concede that measuring wages and hours is manageable, albeit some downward distortion of reported wages by some small percentage figure below Bureau of Labor Statistics levels may occur. Others (for example, Norman Ture) find dire problems in compensation identification/measurement and the arithmetic involved in the Wallich-Weintraub plan, including the costing of the diverse elements of a pay package. The credibility of such criticism is subject to some discount when it is coupled with an almost ideological opposition to incomes policy in any guise.

VII. Concluding Observations

The contention that a *TIP* involves the so-called excess profits tax (*EPT*) problem, that is, measurement of excess over a specified base level, is not fully applicable to *TIP* since *TIP* relies upon a moving, self-updating base—the prior year. This ameliorates, if it does not obviate, the base period abnormalities that have historically plagued *EPT* design.

A *TIP* in the form of a penalty tax on excess wages applicable to large corporations would be quite workable. More elaborate and hybrid forms call for specific appraisal. Both technical and economic issues need further study, but the real obstacle to *TIP* in the form described is not technical infeasibility but one of opposition by conventional economic interest groups concerned that their stakes in the inflation game may suffer in the process of stopping it.

A *TIP* would not cope with all aspects of the complex inflation problem. It would not remove the need for proper restraint in monetary and fiscal policies. It would permit these policies to be used with less frustration and without departure from the free economy.

REFERENCES

Arthur M. Okun and George L. Perry, *Brookings Papers on Economic Activity 2*, Special Issue: *Innovative Policies to Slow Inflation*, Washington 1978.

R. E. Slitor, "Tax-Based Incomes Policy: Technical and Administrative Aspects," a report prepared for the Board of Governors, Fed. Reserve System, Mar. 20, 1978.

N. B. Ture, "Tax-Based Incomes Policy: Pain or Pleasure in Pursuit of Price Level Stability," *Tax Foundation's Tax Rev.*, June 1978, *39*, 23–30.

H. C. Wallich and S. Weintraub, "A Tax-Based Incomes Policy," *J. Econ. Issues*, June 1971, *5*, 1–19.

Welfare Comparisons and Equivalence Scales

By ROBERT A. POLLAK AND TERENCE J. WALES*

Equivalence scales are used in both demand and welfare analysis. In demand analysis they permit us to pool data from households of different sizes, or, more generally, with different demographic profiles. In welfare analysis, they enable us to compare the well-being of such households, since they purport to answer questions of the form: "What expenditure level would make a family with three children as well off as it would be with two children and $12,000?" Such welfare comparisons are generally thought to provide the rationale for different treatments of different family types in income tax or family allowance schedues, or in income maintenance programs.

In this paper we argue that the equivalence scales required for welfare comparisons are logically distinct from those which arise in demand analysis. The usual practice is to base welfare comparisons on equivalence scales estimated from observed differences in the consumption patterns of households with different numbers of children. This is illegitimate. The expenditure level required to make a three-child family as well off as it would be with two children and $12,000 depends on how the family feels about children. Observed differences in the consumption patterns of two- and three-child families cannot even tell us whether the third child is regarded as a blessing or a curse.

In Section I we discuss the type of equivalence scale appropriate for demand analysis (conditional equivalence scales) and, in Section II, the type required to make welfare comparisons (unconditional equivalence scales). Conditional equivalence scales can be estimated from observed differences in the consumption patterns of households with different demographic profiles, but construction of unconditional equivalence scales requires more information than is contained in household consumption data. In Section III we discuss identification of a family's unconditional equivalence scale, while in Section IV we discuss the interpretation of welfare comparisons when families have different tastes. In Section V, the concluding section, we summarize our discussion of welfare comparisons of families with different demographic profiles and question whether such comparisons account for the widespread belief that different treatments of different family types are appropriate.

I. Demographic Variables in Demand Analysis: Conditional Equivalence Scales

In demand analysis, the "objects of choice" are consumption vectors X, and preferences over them depend on an assumed predetermined vector of demographic variables η; we call such a preference ordering "conditional." We denote the conditional preference ordering by $R(\eta)$ and interpret the statement "$X^a R(\eta^*) X^b$" to mean that the family finds X^a at least as good as X^b when its demographic profile is given by η^*. If each family takes its demographic profile as fixed when choosing its consumption pattern, demand analysis need never ask how it would choose between alternatives which differ with respect to the demographic variables; hence, conditional preferences are an appropriate foundation for demand analysis.[1]

*University of Pennsylvania and University of British Columbia, respectively. Pollak's research was supported in part by the National Science Foundation, the U.S. Bureau of Labor Statistics, and the National Institutes of Health.

[1]By "family preferences" we mean the preferences of the adults in the family; preferences of children are ignored. For a family containing one adult, this notion of

In our 1978b paper, we examine a number of alternative ways to incorporate demographic variables into demand analysis by allowing some of the parameters of a demand system to be functions of demographic variables. These functions, which we call conditional equivalence scales, are usually estimated along with the parameters of the demand system by combining data from households with different demographic profiles. The alternatives to this procedure are (i) to analyze separately data from households with distinct demographic profiles or (ii) to combine data from households with different demographic profiles using conditional equivalence scales estimated from other data or specified a priori. The assumption that demand functions are independent of demographic variables, or that per capita consumption of each good is a function of per capita total expenditure, are examples of a priori specifications of conditional equivalence scales.

II. Demographic Variables in Welfare Analysis: Unconditional Equivalence Scales

In contrast to demand analysis, welfare analysis must compare the well-being of a family in alternative situations which differ with respect to its demographic profile as well as its consumption pattern. For example, we might ask whether a family with given tastes would prefer to have two children and \$12,000 or three children and \$13,000 at a particular set of goods' prices. The traditional approach to welfare comparisons ignores the fact that such comparisons cannot be based on conditional preferences but requires a conceptual framework in which preferences are defined over family size as well as goods. In general, welfare analysis requires us to define the objects of choice to include not only the consumption vector, but also the demographic variables, which from the standpoint of conditional preferences are predetermined. We call an ordering over such an augmented set of alternatives an "unconditional preference ordering," and denote it by R: the statement "$(X^a, \eta^a)\ R\ (X^b, \eta^b)$" means that the family finds (X^a, η^a) at least as good as (X^b, η^b). The additional information contained in the unconditional preference ordering is irrelevant for demand analysis, but for welfare comparisons, it is indispensable.

Unconditional equivalence scales are index numbers which reflect the ratio of the expenditures required to attain a particular indifference curve under alternative demographic profiles.[2] Corresponding to the unconditional preference ordering R, we define the "unconditional expenditure function," $E[P, \eta, (P^o, \mu^o, \eta^o)]$, whose value is the minimum expenditure required to reach the indifference curve attained in the price-expenditure-demographic situation (P^o, μ^o, η^o), when the household faces prices P with the demographic profile η. The unconditional equivalence scale $I[(P^a, \eta^a), (P^b, \eta^b), (P^o, \mu^o, \eta^o)]$, is defined by

$$(1)\quad I[(P^a, \eta^a), (P^b, \eta^b), (P^o, \mu^o, \eta^o)] = \frac{E[P^a, \eta^a, (P^o, \mu^o, \eta^o)]}{E[P^b, \eta^b, (P^o, \mu^o, \eta^o)]}$$

If we let the base indifference curve correspond to the "reference situation" (P^b, μ^b, η^b), then the denominator is μ^b and the index is equal to $E[P^a, \eta^a, (P^b, \mu^b, \eta^b)]/\mu^b$. In our

family preferences is unambiguous, but for a family containing two adults, there is an aggregation problem unless the adults' preferences happen to coincide. We ignore this and assume that the notion of family preferences is well-defined. Basing welfare comparisons on family preferences means that we can only compare demographic profiles whose adult compositions are identical; thus, we cannot compare the welfare of a family consisting of one adult and two children with that of a family of two adults and two children.

[2]This corresponds to John Muellbauer's definition of equivalence scales as "budget deflators which are used to calculate the relative amounts of money two different types of households require in order to reach the same standard of living." However Muellbauer uses what we have called conditional equivalence scales to make welfare comparisons, and his paper provides numerous references to other studies which do so. We contend that such an approach is not valid because unconditional equivalence scales rather than conditional equivalence scales are required for welfare comparisons. Our objection to the use of conditional equivalence scales in welfare analysis does not depend on whether families can or do regulate their fertility.

example, such an index would show the percentage expenditure adjustment which would enable a family with three children to attain the same indifference curve it would attain with two children and \$12,000.[3]

We illustrate this with an unconditional preference ordering which is consistent with the familiar linear expenditure system (*LES*) conditional demand functions. Consider the direct utility function

$$(2)\quad W(X,\eta) = \prod_{k=1}^{n} (x_k - b_k^* - \beta_k\eta)^{a_k} + \phi(\eta); \qquad \sum a_k = 1, \quad x_i - b_i^* - \beta_i\eta > 0$$

where η is the number of children in the family; in some very informal sense the function $\phi(\eta)$ represents the "direct" contribution of children to family utility. Substituting the conditional demand functions into this direct utility function yields a "mixed" indirect utility function whose arguments are P, μ, and η:

$$(3)\quad V(P,\mu,\eta) = (\mu - \sum p_k b_k^* - \eta\sum p_k\beta_k) \cdot \Pi(p_k)^{-a_k}(a_k)^{a_k} + \phi(\eta)$$

Solving for μ yields the unconditional expenditure function

$$(4)\quad E(P,\mu,\eta,s_o) = \sum p_k b_k^* + \eta\sum p_k\beta_k + [s_o - \phi(\eta)]\Pi(p_k)^{a_k}(a_k)^{-a_k}$$

where s_o is the value of the utility function (2) evaluated at any point on the base indifference curve. To find the unconditional equivalence scale evaluated at the base indifference curve corresponding to (P^b, μ^b, η^b), we divide the unconditional expenditure function evaluated at $s_o = V(P^b, \mu^b, \eta^b)$ by μ^b. This yields[4]

$$(5)\quad I[(P^a,\eta^a),(P^b,\eta^b),(P^b,\mu^b,\eta^b)] = \{\sum p_k^a b_k^* + \eta\sum p_k^a\beta_k + [(\mu^b - \sum p_k^b b_k^* - \eta^b\sum p_k^b\beta_k)\Pi(p_k^b)^{-a_k}(a_k)^{a_k} + \phi(\eta^b) - \phi(\eta^a)]\Pi(p_k^a)^{a_k}(a_k)^{-a_k}\}/\mu^b$$

III. Identification of Unconditional Equivalence Scales

Since the unconditional equivalence scale corresponding to $W(X,\eta)$ depends on the function $\phi(\eta)$, we must estimate this function. But if we interpret our data in terms of conditional choices (i.e., choices in which the number of children is taken to be fixed or predetermined) the function $\phi(\eta)$ is not identified.[5] All functions $\phi(\eta)$ imply the same conditional demand functions for goods, so information about how a family would reallocate its expenditure among consumption categories as the number of children varies is not sufficient to identify $\phi(\eta)$.

Of course if $\phi(\eta)$ were assumed to be a constant then it would not appear in (5) and the unconditional equivalence scale corresponding to $W(X,\eta)$ could be identified from conditional choices. This appears to be the assumption generally made, although not explicitly, in the literature on equivalence

[3]The conventional cost-of-living index holds the demographic profile fixed and compares the expenditure required to attain a particular indifference curve under alternative price regimes. Such an index can be interpreted as a "subindex" of the unconditional equivalence scale which is itself the "complete index." Pollak develops the theory of subindexes of the cost-of-living index. In this paper we are concerned with complete indexes, or at least with indexes complete enough to include the demographic variables. Subindexes (i.e., conventional cost of living indexes) can be constructed separately for each family type, but such indexes do not permit comparisons of families of different types.

[4]Notice that the unconditional preference ordering corresponding to the direct utility function $\hat{W}(X,\eta) = \Sigma a_k \log(x_k - b_k^* - \beta_k\eta) + \phi(\eta)$ is not the same as that corresponding to $W(X,\eta)$ and hence these two unconditional preference orderings yield distinct unconditional equivalence scales. However, both imply the same *LES* conditional demand functions and, hence, the same conditional equivalence scales.

[5]Whether a particular demographic variable should be treated as predetermined or an object of choice is not automatically resolved by the fact that the variable in question is controlled or chosen by the family, and, hence, could legitimately be treated as an object of choice. For purposes of demand analysis, it is useful to treat family size as predetermined and work with conditional demand functions. When we treat such choices as unconditional, estimation of (unconditional) preferences requires us to reconstruct the feasible set from which the choice was made. But estimation of unconditional preferences is a secondary issue for us. We are primarily concerned with drawing the distinction between conditional and unconditional preferences and arguing that the latter are required for welfare comparisons.

scales and welfare comparisons. However, this assumption has grossly implausible implications for unconditional preferences and unconditional choices involving family size. In particular, consider a "perfect contraceptive society"—one in which there are no economic costs or preference drawbacks associated with fertility regulation. If $\phi(\eta)$ is a constant and $\Sigma p_k \beta_k$ is positive, then the family will have no children; if $\phi(\eta)$ is a constant and $\Sigma p_k \beta_k$ is negative, the family will have as many children as it can. This follows immediately from the fact that when $\phi(\eta)$ is a constant the utility function (3) depends linearly on η.

Another illustration of the counterintuitive results that may occur when we make the transition from household consumption patterns to welfare conclusions by assuming that $\phi(\eta)$ is a constant is provided by the linear expenditure system estimated by the authors (1978a) using *U.K.* household budget data. The estimated conditional demand functions exhibit reasonable price and expenditure elasticities, and reasonable consumption responses to changes in family size. The estimated β's, however, are all negative so $\Sigma p_k \beta_k < 0$. Hence, when $\phi(\eta)$ is assumed to be constant, the unconditional expenditure function decreases with η, and the corresponding unconditional equivalence scale implies that large families need less money than small families to attain any fixed indifference curve.

If unconditional preferences cannot be recovered from conditional demand functions, how can they be discovered? For some demographic variables information about unconditional preferences is revealed by observable choice behavior. For example, in advanced industrial societies where deliberate choice of completed family size is the rule rather than the exception, an argument can be made for treating the observed consumption-family size configurations as observable unconditional choices, using them to infer unconditional preferences, and using these preferences to make welfare comparisons. Thus, in a perfect contraceptive society, if a family chooses to have three children and \$12,000 when it could have had two children and \$12,000, then a revealed preference argument implies that the family prefers the alternative it chose.[6] Other demographic variables (say, race) are not susceptible to deliberate control, while still others (say, the sex of a family's first child) may be moving from the uncontrollable to the controllable category. Unconditional preferences for demographic variables might also be obtained by analyzing responses to direct questions about preferences or hypothetical choices, although economists have traditionally been suspicious of this approach.[7]

IV. Welfare Comparisons with Taste Differences

Taste differences—that is, differences in families' unconditional preferences—substantially complicate welfare comparisons. There are two approaches. The first is to select a particular unconditional preference ordering as the appropriate base for welfare comparisons and proceed as before. The selection is trivial if a particular preference ordering is obviously appropriate as when all families have identical unconditional preferences. It is especially troublesome when systematic differences in preferences are associated with systematic differences in the demographic variables, as is the case with family size or other demographic variables over which families exercise partial or complete control. For some demographic variables it may be plausible to assume that families with different demographic profiles have the same unconditional preferences, or more precisely, that the distribution of unconditional preferences is independent of the distribution of demographic characteristics. But for demographic variables over which families exercise some deliberate control, this independence assumption is clearly unwarranted.

Suppose, for example, that some families

[6]Multiple births create special problems which we ignore. We interpret the \$12,000 as total expenditure on goods and ignore both the labor-leisure choice and the dependence of taxes on demographic variables.

[7]For an example of an equivalence scale constructed from responses to a questionnaire asking individuals what income level corresponds to such verbal evaluations as "good," "sufficient," "bad," etc., see Arie Kapteyn and Bernard van Praag.

have a strong desire for children while others have a weak desire for children. Then the expenditure required to make a family with three children as well off as it would be with two children and $12,000 depends on which unconditional preference ordering it has. Hence, the unconditional equivalence scale depends on which of the two unconditional preference orderings we select as the base. But neither selection compares the welfare levels of families with different tastes. Instead, they compare two situations (for example, three children, $13,000 vs. two children, $12,000) on the basis of a particular preference ordering—whichever one selected is the appropriate base for the comparison.

The second approach to welfare comparisons requires interpersonal (interfamily) comparisons of happiness or satisfaction. Technically, we need a mapping which associates with each indifference curve from one unconditional indifference map a corresponding curve on the other, so that the corresponding curves represent the same levels of happiness or satisfaction. Only if such a correspondence exists can we compare the welfare of families with different tastes in alternative situations—for example, strong desire for children, three children, $13,000 vs. weak desire for children, two children, $12,000.

V. Conclusion

The implications of our analysis of welfare comparisons and equivalence scales should be stated explicitly: 1) Even if all families have identical unconditional preferences, conditional equivalence scales estimated from observed differences in the consumption patterns of families with different demograpic profiles cannot be used to make welfare comparisons; for example, we cannot use such data to determine the amount needed to make families with three children as well off as those with two children and $12,000. Unconditional equivalence scales are required to make welfare comparisons. 2) If tastes vary systematically with demographic characteristics, then the construction of unconditional equivalence scales requires the selection of an appropriate base unconditional preference ordering; theory offers little guidance in making this selection, but there is no selection which permits us to compare the welfare of a family with a strong desire for children with that of one with a weak desire for children. Such comparisons require interpersonal or interfamily comparisons of welfare levels. The question of whether such comparisons are meaningful, and if so, how they can be made, is beyond the scope of this paper.

Our analysis suggests that it is very difficult to make welfare comparisons between families with different demographic profiles. But are comparisons of this sort the principal basis of the widespread belief that it is appropriate to treat different family types differently in income tax or family allowance schedules or in income maintenance programs? We think not. For example, differences in treatment might be justified in terms of effects on the children's present or future welfare, the effects on the children's future productivity, or the effect on the family's fertility.[8] Our analysis implies that differences in treatment cannot easily be justified by an appeal to equity or fairness if this is interpreted in terms of "family preferences" (i.e., the welfare of the adult members of the family). But the arguments one would advance to justify providing children in large families with consumption levels which society somehow establishes to be "socially adequate" are very different from those one would advance for making the adults in large and small families equally well-off. The problem of defining socially adequate consumption levels is a difficult one which has received virtually no attention from economists, in part because of the profession's unfortunate preoccupation with welfare comparisons and equivalence scales.[9]

[8]The relevance of these considerations and of welfare comparisons will vary from one policy question to another

[9]There is no reason to think that conditional equivalence scales have any role to play in the determination of socially adequate consumption levels.

REFERENCES

A. Kapteyn and B. van Praag, "A New Approach to the Construction of Family Equivalence

Scales," *Euro. Econ. Rev.*, May 1976, *7*, 313–35.

J. **Muellbauer,** "Testing the Barten Model of Household Composition Effects and the Cost of Children," *Econ. J.*, Sept. 1977, *87*, 460–87.

R. A. **Pollak,** "Subindexes in the Cost of Living Index," *Int. Econ. Rev.*, Feb. 1975, *16*, 135–50.

——— **and T. J. Wales,** (1978a) "Estimation of Complete Demand Systems from Household Budget Data: The Linear and Quadratic Expenditure Systems," *Amer. Econ. Rev.*, June 1978, *68*, 348–59.

——— **and** ———, (1978b) "Demographic Variables in Demand Analysis," disc. paper no. 78–48, Univ. British Columbia, Dec. 1978.

Comparing Households with Different Structures: The Problem of Equity

By MARILYN E. MANSER*

Design of equitable tax and transfer schemes and analyses of the distribution of well-being among members of society are complicated by the fact that households differ in structure. During the 1970's, a major reexamination of the questions of optimal income tax progressivity and the horizontal equity of taxes has taken place following the seminal work of James Mirrlees; these analyses differ from earlier treatments of tax equity because of the recognition that individuals may value leisure time as well as income. This reexamination of tax equity has for the most part assumed a world of identically structured households with only one potential worker. The present paper surveys issues in tax equity and analyzes them allowing for one- and two-adult households which may differ in their work behavior.

I. Equity in Taxation of Two-Adult Households

The concept of equity in taxation involves two interrelated criteria, horizontal equity and vertical equity. The former requires equal tax treatment of taxpaying units in equal positions, while the latter requires that taxpaying units in unequal pretax positions should pay different amounts of tax.

Before proceeding to discuss alternative measures of position for two-adult households, however, it is necessary to choose the appropriate unit for comparison of well-being. Regardless of whether the individual or the household is taken to be the appropriate unit, it is clear that it is not appropriate to attempt to draw conclusions from an analysis based on the assumption that each individual maximizes his own utility function (defined only over own consumption of private goods) subject to the own budget constraint, since to do so ignores the sharing of goods, tradeoff between leisure times, and pooling of resources which are generally taken to characterize two-person households.

Suppose each household h is composed of two individuals F and M, who value their leisure times x_{1h} and x_{2h}, respectively, as well as their consumption of an aggregate market good x_{3h}. The standard approach to deriving demands or making welfare comparisons for such households assumes the existence of a neoclassical household utility function

$$(1) \qquad U^h = U^h(x_{1h}, x_{2h}, x_{3h})$$

which is maximized subject to time constraints $x_{ih} \leq K$, $(i = 1,2)$ and to the budget constraint

$$(2) \qquad P_{1h}x_{1h} + P_{2h}x_{2h} + x_{3h} = I_h$$

where $I_h = (P_{1h} + P_{2h})K + A_h$ is household full income, A_h is unearned income, and the price of x_3 is taken to be unity.[1]

If the utility functions of the household members differ, it may not be acceptable to force them into this aggregate framework. One alternative which allows for the reconciliation of individual preferences is provided by the bargaining approach which Murray Brown and I have used; it assumes that the individuals in household h maintain their single state utility functions U^F and U^M, agree to a particular bargaining rule, and pool all income. Each of the bargaining models we

*Senior economist, Mathematica Policy Research, Inc. This paper was written while I was at the State University of New York-Buffalo.

[1]The analysis in this paper does not allow explicitly for household production. For example, comparing two households with identical incomes and male labor supplies, the household with more female nonwork hours is clearly better off, since it can produce more commodities and/or can consume more leisure than can the other household. Clearly, if production is occurring in the household, but leisure is valued for its own sake, the distortion of equity and efficiency entailed by income taxation does not arise solely because of the failure to tax nonmarket production.

analyzed can be structured as a problem in which an objective function, which depends on the individual utility functions, is maximized subject to a constraint set, but the objective function cannot be interpreted as "household utility." For each of the rules analyzed, a unique solution X^* is obtained for the household's allocation and distribution problem.

The bargaining approach permits comparisons of well-being between either individuals or households, although if the bargaining process differs between households the latter comparisons must be undertaken based, in some manner, on the utility possibility sets because there is no meaning attached to a comparison of levels of different objective functions. It is not clear that members of two otherwise identical households which differ in the distribution of well-being among the members because of use of different bargaining rules should be taken as being in different positions for the purpose of tax policy; consequently, the household is taken as the unit of comparison here. Clearly an individualistic approach must be adopted for comparison of one- and two-person households.

Returning to the question of equity in taxation, there has been much discussion of the choice between income and consumption as the measure of position for analyses of tax equity. Actual income, properly defined, is commonly accepted as the appropriate measure; see, for example, Henry Simons or Richard Musgrave (1959). Accordingly, Musgrave and Peggy Musgrave argue that a household in which both individuals earn $10,000 is in the same position as another household in which one individual earns $20,000 and the other supplies no labor, although, even if tastes are identical, they may not receive equal utility.

Suggestions for making welfare comparisons and considering tax equity on a basis other than income have appeared during the 1970's. Irwin Garfinkel and Robert Haveman propose a measure of the well-being of households, "earning capacity" (*EC*), which is based on a family's ability to generate income if it uses its human and physical capital at "capacity"; not only the composition of the poverty population but also estimates of effectiveness of transfer programs are sensitive to the choice between *EC* and income as a measure of poverty. Clair Vickery argues that, if the minimum consumption level for the poor requires home production as well as income, "... to base the benefit schedule of an income-support program on an index that defines poverty in terms of money income alone is to create gross inequities across households that vary in their number of adult hours" (p. 27–28). A similar point emerges from work on optimal income taxation. Recognition that individuals value leisure time as well as income has led to a redefinition of horizontal equity to require that two individuals who initially obtain equal utilities should obtain equal utilities after a tax is imposed or changed; see Martin Feldstein (1976) and Musgrave and Musgrave.

Even for a world of individuals, this utility-based definition of horizontal equity is satisfied by income taxes only under very special conditions. Musgrave (1976, p. 7) argues that for individuals with identical preferences but facing differing relative prices, full income is an operational measure of options which is clearly superior to income as a measure of position. If preferences differ, an income tax imposes a lesser burden on "leisure lovers." Although a tax on full income (*FYT*) also will not satisfy the utility-based definition of horizontal equity if preferences differ, Musgrave (1976) recommends full income as an index of horizontal equity for the case where relative prices are the same, since it reflects available options.

The remainder of this section analyses the effects of income taxation and of a *FYT* on equity under the assumption that a utility function (1) exists.[2] (The conclusions will not in general be changed in the more complex case where a bargaining model describes the household decision-making process.) A straightforward extension of the utility-based definition of horizontal equity to the case of

[2]Michael Boskin has noted that efficiency requires taxing male (earned) income at a higher rate than female (earned) income because of differing elasticities of labor supply.

households requires that households who obtain equal utilities in the pretax situation should obtain equal utilities after the tax is imposed.[3] Consider two households, A and B, who possess identical preferences, face wages $P_{iA} \neq P_{iB}$ $(i = 1,2)$, but, in the initial (no tax) situation, attain utility levels, which, in terms of the household indirect utility functions, are

$$(3) \quad V^A(P_{1A},P_{2A},(P_{1A} + P_{2A})K) = V^B(P_{1B},P_{2B},(P_{1B} + P_{2B})K)$$

Even though no unearned income is received, satisfaction of (3) does not require that earned income or full income be equal for A and B. Thus, because relative prices may differ, in general neither a tax on income nor a tax on full income will preserve horizontal equity, although for the special case of homothetic preferences, a proportional FYT will satisfy horizontal equity. But full income again seems preferable to income as a measure of position for it is independent of the labor-leisure choice.

Unless society prefers inequality, vertical equity requires that units in "better" positions pay higher taxes. Analyses of optimal taxation, following Mirrlees, address how taxes should vary by maximizing a Bergsonian social welfare function (SWF), defined over the (identical) utility functions of individuals who value leisure and income, subject to the constraint that a fixed amount of government revenue R be raised for nonredistributional purposes. The initially unexpected result that the marginal income tax rate and the amount of progression may be quite low at the optimum, even under SWF specifications which are very equality preferring, arises for equity as well as efficiency reasons; see Efraim Sadka. The higher the elasticity of substitution between leisure and consumption and, in general, the higher R, the less progressive the tax structure; see, for example, Feldstein (1973). The general conclusions of these studies based on individual preferences would be preserved by studies based instead on household preferences. However, simulations allowing male labor supply to be less elastic than female labor supply, as consistent with most empirical results, would imply less progression than studies based on individual preferences and the low substitution elasticities consistent with observed male labor supply behavior.

For given societal preferences, taxing full income rather than actual income is expected to lead to more redistribution at the optimum due to the avoidance of efficiency loss. But considering taxation in the context of households points out the limited potential of full income as an equitable and operational tax base.[4] Since the potential wage rate for nonworkers is not observed, it would have to be estimated for imposition of a tax on full income if the equity claim—that it represents available options—were to hold. Such a procedure would represent a major departure from the usual practice of levying taxes on observables. Also, if the actual wage rate were always used as a measure of options facing workers, incentives for one household member to supply a small amount of labor at less than the true potential wage would exist for certain price regimes. However, full income merits attention for use instead of actual income as a measure for evaluating the distribution of well-being among equal sized households.

II. Equity in Taxation of Married vs. Single Individuals

Without knowledge of how the household objective function is related to the utility functions of the individuals forming the household, it is not clear how welfare comparisons can be made between married and single individuals. Here, comparisons of well-being between individuals in one- and two-person households are undertaken in a manner

[3]Discussions of equity necessarily involve interpersonal utility comparisons. Anthony Atkinson and Joseph Stiglitz state that "When individuals have the same indifference curves, it is natural simply to use the same cardinal number of the indifference curves for different individuals, but if tastes differ, this is no longer so" (p. 71). See also Musgrave (1976).

[4]Similar problems would arise if the graded earning tax proposed by Jonathan Kesselman were implemented in a world of households

consistent with the bargaining analysis of household decision making mentioned earlier but without postulating a particular bargaining rule. The analysis is based on the assumptions that all females possess identical utility functions as do all males, and that there is no "caring." Further, both one- and two-person households are assumed to consume a market good x_3, which is a "pure public" or "shared" good to the household, that is, one spouse's consumption of x_3 does not reduce the amount available to the other, and neither is excluded from consuming x_3.

If single, two individuals obtain utilities $U_0^F = U^F(x_{10}, P_1(K - x_{10}))$ and $U_0^M = U^M(x_{20}, P_2(K - x_{20}))$. If they marry, then the single state consumption bundle is still attainable, although not optimal, in general. If it is purchased it will provide them with utilities $U^F = U^F(x_{10}, x_{30})$ and $U^M = U^M(x_{20}, x_{30})$, where $x_{30} = P_1(K - x_{10}) + P_2(K - x_{20})$, which exceed the corresponding single state utilities. Thus, given the existence of this shared good, an individualistic utility-based approach to equity requires that, if two individuals marry, their taxes be increased (or transfer payments be reduced). This conclusion is preserved if private goods are also consumed so long as some goods are shared. Economies of scale due to lower per unit cost of purchasing goods in greater quantity also may provide consumption gains to households. Clearly, use of an imperfect proxy (i.e., legal marriages) to identify those individuals gaining from household formation for the purpose of levying taxes or determining transfer payments does cause equity distortions.

It is frequently argued that a single individual should pay higher taxes than a married couple with equal income. Although this is in accord with the utility-based approach to equity if none of the individuals being compared provide any labor and if not all goods are shared goods to households, it may not be if leisure times differ. To see this, consider a single individual Ms. B, who consumes less leisure than Ms. A, who is part of a household with the same income; under the above assumptions, $x_{1B}^* < x_{1A}^*$ implies $U^F(x_{1B}^*, x_{3B}^*) < U^F(x_{1A}^*, x_{3A}^*)$, since $x_{3B}^* = P_{1B}(K - x_{1B}^*) = x_{3A}^* = P_{1A}(K - x_{1A}^*) + P_{2A}(K - x_{2A}^*)$.[5] If some private goods were also consumed, Ms. A's utility could still exceed that of Ms. B.

Even if individuals possess the same tastes, without knowledge of the utility functions it is not possible to compare the well-being of married and single individuals. The problem with the use of actual income was discussed above. Full income is not an appropriate index for such a comparison either, since a single individual would be better off than a household with the same full income. Nor is per capita full income appropriate, since it does not reflect the consumption gains accruing to two-person households.

III. Concluding Remarks

This paper argues that household full income is preferable to earned income as an index for comparing the well-being of households of equal size because it provides a measure of available opportunities which is independent of labor-leisure choice. However, no measure which is independent of preferences is adequate for comparing the well-being of members of one- and two-adult households.

The point that income taxation in general violates the utility-based definition of tax equity is not a new one. However, considering the problem of tax equity in the context of households suggests that the distortion of equity caused by failure to tax adult nonwork hours is not insignificant, since a large part of the variation in nonwork hours of two-adult households arises because many women provide little or no market labor, while most of the others work close to full time. Introduction of a tax deduction based on a summary measure of work effort, say on the number of full-time workers in the household, could perhaps improve equity.

An extremely important problem not addressed above involves comparison of well-being of households with different numbers of

[5]For the simple utility specifications considered explicitly here, F must provide some earned income in order for M to attain a utility level which exceeds that which he would obtain if single. Clearly, that is not required if there is caring or household production.

children. Even if it is assumed the households have identical preferences, the utility-based approach to equity does not lead to any general conclusions regarding which of two households with different numbers of children but equal incomes and adult work hours is better off. This occurs because of offsetting effects; although expenditures for "private goods" consumed by children and time spent caring for them may reduce utility to parents, children provide a direct source of utility to parents and, at least in developed countries, represent a consumption choice by adults.

REFERENCES

A. B. Atkinson and J. E. Stiglitz, "The Design of Tax Structure: Direct vs. Indirect Taxation," *J. Publ. Econ.*, July/Aug. 1976, *6*, 55–75.

M. J. Boskin, "Efficiency Aspects of the Differential Tax Treatment of Market and Household Economic Activity," *J. Publ. Econ.*, Feb. 1975, *4*, 1–25.

M. Feldstein, "On the Optimal Progressivity of the Income Tax," *J. Publ. Econ.*, Nov. 1973, *2*, 357–76.

———, "On the Theory of Tax Reform," *J. Publ. Econ.*, July/Aug. 1976, *6*, 77–104.

I. Garfinkel and R. Haveman, "Earnings Capacity and the Target Efficiency of Alternative Transfer Programs," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 196–204.

J. R. Kesselman, "Egalitarianism of Earnings and Income Taxes," *J. Publ. Econ.*, Apr./May 1976, *5*, 285–301.

M. Manser and M. Brown, "Marriage and Household Decision-making: A Bargaining Analysis," *Int. Econ. Rev.*, forthcoming.

J. A. Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, 195–208.

Richard A. Musgrave, *The Theory of Public Finance*, New York 1959.

———, "ET, OT, and SBT," *J. Publ. Econ.*, July/Aug. 1976, *6*, 3–16.

——— **and Peggy B. Musgrave,** *Public Finance in Theory and Practice*, 2d ed., New York 1976.

E. Sadka, "On Progressive Income Taxation," *Amer. Econ. Rev.*, Dec. 1976, *66*, 931–35.

Henry C. Simons, *Personal Income Taxation*, Chicago 1938.

C. Vickery, "The Time Poor: A New Look at Poverty," *J. Hum. Resources*, Winter 1977, *12*, 27–48.

The Social Security Benefit Structure: Equity Considerations of the Family as its Basis

By CAROL T. F. BENNETT*

The secondary benefit structure of the Old Age, Survivors, and Disability Insurance system (*OASDI*) transfers $25 billion per year to families of retired, deceased, and disabled workers without appreciable regard to past contributions or to need. Since these benefits are financed through payroll taxes, the insurance premiums of larger families are subsidized by the contributions of individuals and smaller families, regardless of ability to pay. This subsidy exists even with the strong weighting of the system in favor of lower income workers. The present research indicates that returns on Social Security contributions vary more by family pattern than by any other variable. Moreover, the enactment of several recent congressional bills would further expand the transfer among family types. Research suggests that if equity among families and individuals is an issue of concern, then alternative means of financing Social Security cost increases should be implemented.

I. The Model

The transfer from one family pattern to another can be measured in terms of the ratio of accumulated (real) benefits to accumulated contributions. For any individual or family:

$$\text{Accumulated benefits} = \sum_{t=1}^{n} B_t(S,R,E_t,L_t,M_t,C_t,P_t,D_t,A_t)\,(1 + r)^{n-t}$$

$$\text{Accumulated taxes} = \sum_{t=1}^{n} T_t(S,R,E_t,L_t,A_t)\,(1 + r)^{n-t}$$

where B_t = expected real benefits received in year t, T_t = expected real taxes paid in year t, S = sex, R = race, E = real earnings, L = labor force participation, M = marital status, C = number of children, P = number of parents supported, D = divorce status, A = retirement age, and r = assumed real rate of interest.

The model used to simulate the Social Security benefit and contribution structure is actuarial and deterministic. It consists of variables specified by alternative family patterns, equations consistent with current Social Security law, and parameters given by economic and demographic data. The dollar transfer to or from one individual or family is equal to the difference between accumulated contributions and accumulated benefits. This transfer is measured in constant dollars over the course of an active Social Security account, assumed to be between ages 20 and 99. The ratio of accumulated benefits to accumulated contributions can be interpreted as the return from Social Security relative to an alternative annuity and insurance amount that could potentially be purchased by a worker's contributions.

Thirteen variables describe family patterns of workers. Family marital status and sex appear as either single male, single female, or married couple. Workers are considered single if never married or if divorced and not remarried by age 50; the demographic parameters provide for the event of widowhood. For married couples, the Social Security accounts of the husband and wife are computed separately, and the resulting expected benefits and contributions are summed annually. Accounts for one or more divorced wives may be computed if there had been at least ten years of marriage. In cases of divorce, a variable specifying marriage duration describes the change in family status.

Additional dependents are described through the child and parent variables. In addition to the total number of children, the number of children born to the current and/or former marriage is specified. This is neces-

*Visiting assistant professor, University of Houston.

sary since benefits are available for nonaged spouses and widow(er)s with children in their care. The timing of first childbirth is a variable, though all subsequent children are assumed three years apart in age. Parents dependent on the worker for half their support also are included in the model, with both male and female workers each potentially supporting a mother and/or father.

Three variables describe labor force participation. A woman is assumed to leave the labor force only once—at the birth of her first child—but the age at her return to paid employment, if ever, is variable. Male workers are expected not to leave the labor force until retirement. Retirement age can take on the values of 62 to 70, though the assumption is made that spouses retire at the same age.

The demographic parameters were drawn from Social Security Administration and National Center for Health Statistics data. The probabilities of living, dying, and being deceased for each sex and single year of age were expanded from mortality tables to include only those alive on their twentieth birthday. All persons were assumed deceased on their hundredth birthday. Additional statistics were used for race differentials and for projected mortality in the year 2000.

Disability probabilities were derived from the *Social Security Bulletin Annual Statistical Supplement, 1975*. For each sex, the actual number of disability beneficiaries by single year of age was divided by the estimated number of living workers insured for disability. This proportion was further discounted for mortality. Child mortality and disability statistics were estimated from these data also, but differentials by race and sex were ignored.

Earnings, inflation, and interest rate parameters were drawn from Department of Commerce statistics. Mean earnings by sex, race, and five-year age group were estimated for 1978 by expanding 1975 amounts by the change in the average wage index. Three levels of earnings were used for each family pattern: the mean level, a higher and a lower transformation of the mean. The interest rate used in the model was the average yield on a portfolio divided equally between long-term U.S. Treasury and corporate bonds over the preceding twenty-year period, after the current rate of inflation had been subtracted annually. This method resulted in a real long-term interest rate estimated at 1.9 percent. Since the 1977 amendments to the Social Security Act compensate for inflation in both the benefit and contribution structures, all amounts denominated in dollars in the model are stated in real terms. Finally, the payroll tax is assumed fully paid by the worker; the results correspond well with this assumption.

With these variables and parameters, the model simulates the expected returns on contributions by multiplying in each year the dollar amounts of benefits and contributions by the probabilities of receiving and paying them. Contributions are a function of earnings by each worker in a family and the tax rate specified by law for each year, up to the taxable maximum amount. Benefits are a function of average earnings (converted into primary insurance amount) for each worker and the number of family members. The equations of the model compute the primary worker's and the secondary family benefits, control for the receipt of more than one benefit, and constrain the total amount to the family maximum benefit in each year.

The estimates presented here tend to be conservative. The ratios are higher if expected future mortality and disability patterns are used, if administrative costs are entered into the benefit side as they are implicitly into the contributions side, if retirement age is assumed to be 63 instead of 65, if husbands are assumed several years older than their wives, if first children are born to parents older than age 25, or if children are born more than three years apart. Finally, use of a lower interest rate raises all ratios significantly, since contributions paid in early years accumulate over a longer period than do benefits received in later years.

II. Empirical Results

The empirical results indicate that for workers entering the *OASDI* system in 1978, returns vary more by family status than by

other variables. Consider, for example, several families with identical total earned income ranging between $12,000 and $19,000 over their worklives. A childless family where the wife is continuously employed pays $77,600 more in accumulated Social Security contributions than it receives in accumulated benefits. If the family has two children, the excess contribution is $42,500; with four children, the excess equals $17,000. A family where the wife is never in the paid labor force transfers about $10,000 into the system if there are no children, but receives a subsidy equal to $12,800 if there are two children and $36,000 if there are four children. A single male receiving the same income would subsidize the system by more than $130,000, while a single female would transfer about $100,000.

Dollar amounts are not directly comparable across earnings levels or in cases where family earnings vary due to discontinuous female labor force participation. The ratios of accumulated benefits to accumulated taxes may be compared, however. Table 1 presents these ratios by family category and earnings level. The mean earnings level refers to mean male and female earnings for year-round, full-time workers: these earnings are thus not constant within each column and ratios can be compared only generally. A two-earner low-income family receives about the same wages as a single-earner, mean income family; a two-earner, mean income family receives about the same income as a high-earning male. Mean female earnings are about 60 percent of mean male earnings. These earnings levels thus roughly approximate socioeconomic class.

Within each family type, ratios are about 40–60 percent lower for the highest than for the lowest earning workers. For example, a single male with mean earnings can expect to receive 46.9 percent of the benefits that his contributions would warrant in an alternative system; a low-earning male can expect to receive 58.6 percent, and a high earner 36.3 percent. Overall, the ratios correspond with internal rates of return varying from −1.4 percent for a high-earning single male to 3.8 percent for a low-earning family with a grandchild.

TABLE 1—RATIOS OF EXPECTED BENEFITS TO EXPECTED CONTRIBUTIONS FOR WORKERS ENTERING THE SOCIAL SECURITY SYSTEM IN 1978, BY FAMILY TYPE AND EARNINGS LEVEL

	Earnings Level		
Family Type	Mean	Low	High
Single worker			
Male	.469	586	.363
Female	.776	.817	.572
Male with 2 children	.612	.761	.473
Female with 2 children	.875	.920	.644
Married couple			
No children, wife with no *lfp*[a]	.957	1.196	741
No children, wife with full *lfp*	.599	693	453
No children, 4 parents, wife with full *lfp*	.716	825	.540
2 children, wife with 35 years *lfp*	788	905	593
2 children, wife with 5 years *lfp*	1.041	1.262	.794
4 children, wife with 35 years *lfp*	881	1.014	.667
4 children, wife with 5 years *lfp*	1.134	1.376	865
1 child, 1 grandchild, wife with 5 years *lfp*	1 392	1.690	1.062
Divorced male worker			
2 children, former wife with 25 years *lfp*	816	944	611
2 children, 2 former wives with 35 years *lfp*	.808	.916	620
Remarried, 4 children, current and former wife each with 2 children and 25 years *lfp*	1.015	1.112	.736

[a]Labor force participation (*lfp*)

Family pattern is even more important when race is considered than it is in the general case. For example, single females with identical earnings can expect to receive ratios of benefits to taxes of .786 if white and .702 if black. If they are supporting two children, however, the relative returns are reversed, with ratios rising to .874 for whites and .897 for blacks. This effect is stronger with more children, since black workers show far higher incidence of both mortality and disability than do white workers.

The advantages of early retirement also vary by family type. A single worker gains only marginally from early retirement due to the actuarial reduction of benefits. The ratio for a family with four children, however, rises to .911 for retirement at age 63, up from .881 at age 65, for an ultimate saving of over $10,000.

Social Security ratios are different where children are raised outside a nuclear family. Several cases involving divorce were examined. For a family with two former wives, one of whom is caring for two children, the ratio of benefits to taxes is .808 if both women have thirty-five years labor force experience. With twenty-five years in the labor force and two children born to each wife, the family ratio may rise to 1.015.

III. Policy Alternatives

Many changes in Social Security policy have been proposed recently in Congress. Two of these would have a major impact on the equity of the system among individuals and families: general revenue financing and combining the earnings accounts of husbands and wives.

If full general revenue financing were to replace the payroll tax for Social Security, annual contributions would be a function of family income and the number of current family members. Presumably, all nonretired families would pay some tax for Social Security purposes, the rate perhaps varying between 7 percent for large low-income families and 16 percent for small high-income families. Moreover, since all income would be taxed, high-earner families would be likely to pay an additional amount on unearned income.

In order to approximate reasonable tax rates, the following formulas are used in the model:

$$\begin{aligned}\text{Tax rate} &= .07 + .10/(1 + \text{family size}) \quad \text{for income less than \$13,000}\\ &= .09 + .10/(1 + \text{family size}) \quad \text{for income \$13,000 to \$30,000}\\ &= .11 + .10/(1 + \text{family size}) \quad \text{for income over \$30,000}\end{aligned}$$

where family size in each year equals the sum of probabilities of life and dependence for all potential family members. Benefits remain a function of earnings alone and not contributions, though new laws and the model could specify this change also.

The results show a significant decrease in equity across family patterns. As a comparison of Table 2 with Table 1 shows, returns to mean income, intermediate size families vary only slightly with this change. However, returns to a couple with no children and full labor force participation fall and the return to a family with four children and minimal female paid employment rises, such that the differential across family patterns increases sharply. When different income levels are compared, the decreases in equity are larger, with large families with low female labor force participation benefitting most from the policy change.

The second policy alternative involves combining the earnings accounts of husbands and wives. In each year, earnings of the husband and wife are added, and one-half of the combined amount is attributed to the Social Security account of each spouse.

This reform provides a higher return to married women participating in the labor force than the small increment they currently receive from their own contributions. It also allows disability coverage for women working in the home and benefits to children from loss of their mothers' services through disability or death. However, for families with low female labor force participation, the ratio falls significantly since widow's benefits remain a set percentage of the worker's lessened primary insurance amount. Concern with the adequacy of benefits would certainly overrule the

TABLE 2—THE EFFECTS OF POLICY ALTERNATIVES ON RATIOS OF EXPECTED BENEFITS TO EXPECTED CONTRIBUTIONS

Family Type	General Revenue Financing Income Level Mean	Low	High	Combined Earnings Accounts Income Level Mean	Low	High
Single worker						
Male	.408	.575	.288	–[b]	–	–
Female	.763	.802	.482	–	–	–
Male with 2 children	.604	.860	.423	–	–	–
Female with 2 children	982	1.033	.608	–	–	–
Married couple						
No children, wife with no *lfp*[a]	954	1.367	.668	.752	.913	.640
No children, wife with full *lfp*	552	.682	.339	.634	.701	.568
No children, 4 parents, wife with full *lfp*	704	.869	.427	.750	.830	.673
2 children, wife with 35 years *lfp*	.786	.991	.488	.809	.916	.607
2 children, wife with 5 years *lfp*	1 078	1.537	721	.883	1.060	.741
4 children, wife with 35 years *lfp*	915	1.158	.568	1.001	1.133	.753
4 children, wife with 5 years *lfp*	1 217	1.745	.812	1.096	1.316	.919
1 child, 1 grandchild, wife with 5 years *lfp*	1 428	2.037	.957	1.143	1.368	.957
Divorced male worker						
2 children, former wife with 25 years *lfp*	815	1 068	.513	832	.974	656
2 children, 2 former wives each with 35 years *lfp*	804	1 064	.521	.781	.899	.602
Remarried, 4 children, current and former wife each with 2 children and 25 years *lfp*	1 031	1 232	.630	1.105	1.212	.831

[a]Labor force participation (*lfp*)
[b]Dash = not applicable

gains in equity resulting from enactment of this policy. Finally, ratios are not raised at all for single parents, many of whom have higher labor force participation and less favorable economic circumstances than do families. If tax rates were raised to finance higher benefits for two-earner families, returns to unmarried workers would be reduced.

IV. Conclusions

There is little rationale in contemporary economic circumstances for allocating Social Security benefits on the basis of family pattern. The Bureau of the Census statistics clearly show that the most affluent male workers are most likely to have two to four children and wives whose main activities are outside the paid labor force. Single males are among the least affluent workers, and single mothers may be the least able to pay high payroll taxes.

An alternative structure would be to incorporate Social Security benefits and contributions into the income tax system. Families with low earnings would receive a tax credit against payroll taxes and low-income beneficiaries would continue to pay no tax on benefits received. Benefits received by the families of higher income workers could then be taxed at the family's marginal tax rate—without the expense of means testing—to compensate for the transfer of social resources in their favor.

REFERENCES

U.S. Bureau of the Census, *Current Population Reports*, Series P-60, No. 105, Washington, June 1977.

U.S. National Center for Health Statistics, *United States Life Tables, 1969–71*, Washington 1975.

U.S. Office of Business Economics, *Surv. Curr. Bus., Biennial Edition 1975*, Washington.

U.S. Social Security Administration, "United States Population Projections for OASDHI Cost Estimates," actuarial study no. 72, Office of the Actuary, Washington 1974.

U.S. Social Security Administration, *Social Security Bulletin Annual Statistical Supplement, 1975*, Tables 50, Washington.

Financial Policies in Open Economies

By DALE W. HENDERSON*

In analyzing the extent to which alternative financial stabilization policies can be expected to dampen the effects of shocks to macro-economic equilibrium in a single open economy, it has often been assumed that the authorities must choose between fixing the exchange rate and allowing it to fluctuate freely. Which of these two pure intervention policies is better usually depends not only on the source of the shocks to the economy (see Robert Mundell, Jerome Stein, and Edward Tower and Thomas Willet), but also on the specification of monetary policy. The nature of the truly optimal financial policy is determined by the kind of information available to the authorities about the structure of the economy and about the shocks to which it is subjected. Under plausible assumptions it is not optimal for a single open economy to adopt either pure intervention policy. However, interactions in a two-country world economy must be considered when choosing financial policies, and an agreement to pursue a pure intervention policy may lead to better outcomes than those implied by noncooperative behavior.

I. Shocks and Financial Policies in a Single Open Economy

The outcomes of alternative financial policies in a single open economy can be illustrated by employing a discrete time model in which asset portfolios are balanced at the beginning of each period.[1] In Figure 1, X_0X_0 is an equilibrium schedule for the single home good which is purchased by both home residents and foreigners; an increase in the home interest rate, which lowers demand, must be accompanied by a decline in home output The line M_0M_0 is an equilibrium schedule for home money which is held by home residents alone; an increase in the interest rate, which reduces money demand, must be offset by a rise in output. The line B_0B_0 is an equilibrium schedule for the single security denominated in home currency which is held by both home residents and foreigners; an increase in the interest rate, which raises demand for the home security, must be accompanied by an increase in output which lowers demand. It is assumed that home money, the home security, and a single security denominated in foreign currency are strict gross substitutes, so the M_0M_0 schedule must be steeper than the B_0B_0 schedule.[2] Either the exchange rate,

*Board of Governors of the Federal Reserve System. Russel Boyer, Ralph Bryant, Peter Clark, Lance Girton, Gene Grossman, C. Michael Jones, and Janet Yellen provided useful comments. I alone am responsible for the remaining shortcomings of the paper. The analysis and conclusions of this paper should not be interpreted as representing the views of the Board of Governors of the Federal Reserve System or anyone else on its staff.

[1]A description of a continuous time version of this model is provided by the author. Ralph Bryant's analysis of the effects of shocks under alternative financial policy regimes is similar to the one of this paper.

[2]This footnote contains a terse description of one specification of the financial sector from which the conclusions in the text can be derived. Home residents hold home money, and home and foreign securities Foreign residents do not hold home money The fraction of home nominal wealth which home residents hold in the form of home money depends positively on home output measured in physical units, and the fractions they hold in home and foreign securities depend negatively on home output The fraction of wealth held in each asset by home residents and foreigners' demand for home securities measured in foreign currency depend on the home interest rate and the foreign interest rate augmented by the expected rate of depreciation of the home currency The expected rate of depreciation of the home currency is an increasing function of the gap between a constant "long-run equilibrium" value of the exchange rate and its current value. The supply of home securities available to private agents is equal to the exogenous supply of fixed-nominal-value, variable-interest-rate bonds issued

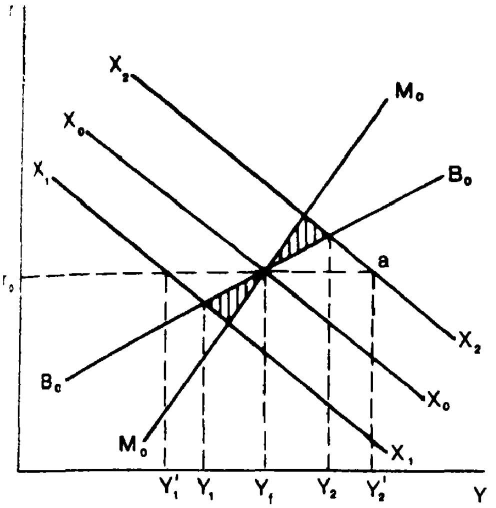


FIGURE 1

defined as the home currency price of foreign currency, or foreign exchange reserves, defined as the home authorities' holding of foreign securities, change in a manner described below until the three schedules have a common intersection point. In Figure 1, X_0X_0, M_0M_0, and B_0B_0 intersect at the "full-employment" level of home output (Y_f). It is supposed that the home currency price of the home good and the foreign currency price of a single foreign good, which is different from the home good, are fixed in the short run and that the foreign authorities act to keep the foreign interest rate and foreign output constant

The home authorities have both home and foreign securities as assets and the money supply as a liability.[3] They can choose as policy instruments and set values for any two of the following four financial variables: the money supply, foreign exchange reserves, the interest rate, and the exchange rate. The values of the other two financial variables are then determined by the model. The authorities conduct financial policy using two kinds of financial market operations: 1) monetary operations, exchanges of home securities for money with private agents; and 2) intervention operations, exchanges of home securities for foreign securities with private agents. Under an "aggregates constant policy" the money supply and foreign exchange reserves are kept unchanged at chosen values, while under a "rates constant policy" monetary and intervention operations are employed to keep the interest rate and the exchange rate constant at selected values.

Consider the effects of stochastic shifts in the XX schedule in the range between X_1X_1 and X_2X_2 shown in Figure 1. These shifts might result from changes in home or foreign saving behavior, or from changes in preferences between home and foreign goods either at home or abroad.[4] If the authorities pursue an aggregates constant policy, levels of output between Y_1 and Y_2 result. For example, suppose an increase in demand for the home good shifts the XX schedule to X_2X_2. Output increases, creating an excess demand for home money and an excess supply of home securities. Under plausible assumptions these disequilibria can be removed only by a rise in the interest rate and an appreciation of the home currency. It is assumed that an appreciation of the home currency raises excess supply in the markets for the home good,

---

by the government of the home country minus the holdings of the authorities.

[3]Values for only two of the three items on the authorities' balance sheet can be chosen independently. If the authorities have the monetary base as a liability, uncertainty in the relationship between the monetary base and the money supply can affect the analysis of financial policy in open economies as explained by Bryant.

[4]In a beginning-of-period-balancing model a change in saving behavior does not affect asset demands No attempt is made here to classify shocks as "real" or "monetary" for two reasons First, these adjectives have been used in different ways by two sets of authors In most of the literature on the analysis of financial policies in closed and open economies real shocks are shifts in the aggregate demand for goods, monetary shocks are shifts in money demand, and aggregate supply adjusts passively to fulfill aggregate demand, so there are no stochastic shifts in aggregate supply. In the contributions of Stanley Fischer and Jacob Frenkel, as in much of the literature on indexation, however, real shocks are shifts in aggregate supply, and monetary shocks are shifts in the quantity-theory money demand function. Second, there is more than one financial market in the system considered here, and of the three kinds of shifts in financial markets considered, only two involve shifts in money demand

home money, and the home security.[5] These assumptions imply that as the home currency appreciates, the X_2X_2, M_0M_0, and B_0B_0 schedules shift toward one another, until they intersect at a point in the shaded triangle above X_0X_0.

If instead the authorities pursue a rates constant policy, levels of output between Y_1' and Y_2' result. If the XX schedule shifts to X_2X_2, then the new equilibrium is at point a. Since there is no change in the exchange rate the XX schedule does not shift from X_2X_2. The MM and BB schedules are shifted to the right by monetary and intervention operations until they pass through point a. An expansionary monetary operation, a purchase of home securities with home money, shifts both MM and BB to the right. However, BB is shifted farther since increases in income raise the demand for money by more than they reduce the demand for home securities because the demand for foreign securities is also reduced. Thus, in order to keep both the exchange rate and the interest rate constant, the authorities must undertake an intervention operation, a sale of home securities in exchange for foreign securities, so that the BB schedule does not shift farther to the right than point a. When the only source of shocks to equilibrium is stochastic shifts in the XX schedule, an aggregates constant policy leads to less variation in output than a rates constant policy.

Exactly the opposite conclusion is reached when stochastic shifts in the BB schedule between B_1B_1 and B_2B_2 shown in Figure 2 are considered. These shifts result from changes in preferences between home and foreign securities either at home or abroad. If the authorities pursue an aggregates constant policy, levels of output between Y_1 and Y_2 result. Suppose a shift in asset preferences toward home securities and away from foreign securities causes the BB schedule to move to B_2B_2. The increase in demand for home securities leads to a decrease in the home interest rate, which in turn causes an excess demand for home money. In order for

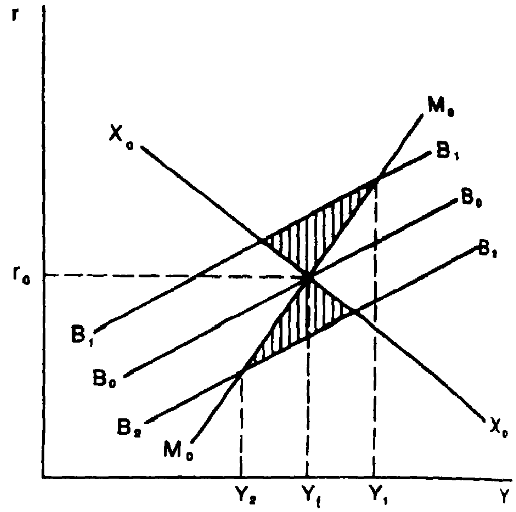


FIGURE 2

equilibrium in the financial markets to be reestablished, the home currency must appreciate. Appreciation causes the three schedules to shift together as before, so the new equilibrium must lie in the shaded triangle below B_0B_0. Output may fall, rise, or remain the same since the changes in financial variables have opposite effects on demand for the home good.

If instead, the authorities pursue a rates constant policy, output definitely remains unchanged. The BB schedule is shifted back to B_0B_0 by an intervention operation consisting of a sale of home securities and purchase of foreign securities. When the only source of shocks to equilibrium is stochastic shifts in the BB schedule, a rates constant policy leads to less variation in output than an aggregates constant policy.

Two other possible sources of stochastic shocks to equilibrium are 1) shifts in home residents' preferences between home money and foreign securities which cause movements in the MM schedule and 2) shifts in home residents' preferences between home money and home securities which cause movements in both the MM and BB schedules. In both these cases, a rates constant policy leads to less variation in output than an aggregates constant policy.

[5] The financial-market assumptions are implications of the specification of the financial sector in fn. 2 and many other plausible specifications.

A familiar conclusion can be confirmed with the diagram if the additional assumption is made that the excess demand for home money does not depend on the exchange rate:[6] when either the money stock or the interest rate is kept constant, 1) with stochastic shocks in the home good market, output variation with a freely floating exchange rate is always less than or equal to that with a fixed exchange rate; and 2) with stochastic shifts in or between the markets for home securities and home money, output variation with a fixed exchange rate is always less than or equal to that with a freely floating exchange rate.

II. Information and Financial Policies in a Single Open Economy

Assume for simplicity that the authorities wish to minimize the expected squared deviations of output from Y_f. How should they proceed when the economy is buffeted by all of the types of shocks considered above? What financial policy is optimal depends on what information the authorities have about the structure of the economy and about the shocks to which it is subjected. Suppose the authorities operate in an environment in which they know, or have unchanging subjective beliefs about, the nonstochastic coefficients of the three linear market equilibrium relations and the joint distribution of the additive stochastic terms. Suppose also that they cannot observe output, and cannot observe or, at least, do not respond to movements in the two financial variables they do not fix when choosing their monetary and intervention policies. In this setting it makes sense to compare alternative pure financial policies. Alternative certainty equivalent policies should be compared, and when there are two policy instruments, as in the system under consideration, one can be set arbitrarily. Then once a policy is found to be superior it is followed period after period unless there are changes in the parameters of the system or the joint distribution of the stochastic disturbances (see William Poole and Benjamin Friedman).

The diagrammatic analysis above suggests one simple kind of conclusion; for example, given the coefficients of the system and all of the other parameters of the joint distribution of the disturbances, there exists a variance of the disturbance term in the market for the home good large enough to insure that an aggregates constant policy leads to lower expected loss than a rates constant policy. Additional conclusions must be based on calculations of expected losses. Suppose that the three equilibrium relations are normalized on income and that the variances of the normalized disturbances are equal. An aggregates constant policy may or may not be better than a rates constant policy whereas under similar assumptions in a closed economy a money supply constant policy dominates an interest rate constant policy (see Poole). An aggregates constant policy is superior (inferior) to a rates constant policy for large values of the degree of substitutability between home and foreign securities (the responsiveness of home good demand to changes in the exchange rate).

The authorities should proceed differently in a second environment in which the only difference is that the coefficients of the model are stochastic variables which have a joint distribution with the additive stochastic terms that is known to the authorities. As before, it is logical to consider alternative pure financial policies. However, in general, certainty equivalent policies are not optimal, and all policy instruments, potentially two in the system considered here, are set at well-defined optimal levels even though there is only one target variable. This is because different values of the same instrument imply different variances for output (see William Brainard). Again, once chosen, instrument levels are not varied. Two questions are of interest: which pure policies should be chosen and how far should the instruments be set from their certainty equivalent levels.

Optimal behavior for the authorities can be

[6] The excess demand for home money would be independent of the exchange rate if the demand for home nominal balances deflated by the price of the home good, instead of a price index which included the exchange rate, depended on only home output measured in physical units and the home interest rate.

described in yet a third environment in which they know the nonstochastic coefficients of the market equilibrium relations and can observe and respond to changes in the two financial variables not chosen as policy instruments. In this setting one policy instrument can be set arbitrarily. The authorities should choose a linear rule which tells them how to set the other policy instrument given the levels of the remaining two financial variables which can be regarded as information variables. In general, the coefficients of the decision rule will be functions of both the coefficients of the model and the parameters of the joint distribution of the additive disturbance terms. While the decision rule is the same period after period, the value of the variable policy instrument is changed from period to period since the authorities can learn something about the shocks in the current period from observations on the two information variables (see Friedman; John Kareken, Thomas Muench, and Neil Wallace; Poole).

Consider the case in which the interest rate and the exchange rate are chosen as policy instruments and the exchange rate is kept fixed. Suppose the authorities make a trial choice of the interest rate which would lead to an output equal to Y_f if there were no disturbances. However, there are disturbances, and, as a result the authorities will observe a money supply and a value of foreign exchange reserves which are different from the ones which would be associated with Y_f if the disturbances were zero. Given the exchange rate, the trial choice of the interest rate, and the observed financial aggregates, the three market relations can be used to eliminate unobservable output and to solve for two linear combinations of the current disturbance terms. These two linear combinations can in turn be employed to form an optimal estimate of the current disturbance term in, for example, the market for the home good, using the parameters of the joint distribution of the disturbance terms. With this optimal estimate of the current disturbance in the market for the home good, the authorities can then choose a new value for the interest rate which assures that the expected value of output is equal to Y_f, given the current disturbances. This new choice of the interest rate will imply new values for the money supply and the authorities' foreign exchange reserves. The differences between the original choice of the interest rate and its final value and between the values of the money supply and foreign exchange reserves implied by the original choice of the interest rate with the disturbance terms set equal to zero and their final implied values can all be expressed as linear functions of the two calculated linear combinations of the disturbance terms, so the required adjustment in the interest rate can be written as a linear function of the deviations in the aggregates. This linear function is the decision rule.[7] When the interest rate is chosen as the variable policy instrument the coefficients on the deviations in the aggregates are both zero if there is no disturbance in the market for the home good.

The implication that one financial policy instrument, the exchange rate in the example above, can be set arbitrarily depends crucially on two assumptions: the assumption that the authorities are concerned only about squared deviations of output from Y_f, and the assumption that the coefficients of the system are known with certainty. If the authorities were also concerned, for example, about squared deviations in interest-sensitive consumption from some desired level, optimal financial policy would involve variations in both policy instruments, so the exchange rate would have to vary no matter whether it was chosen as a policy instrument or was used as an information variable. Likewise, if the coefficients of the model were stochastic variables all financial variables, including the exchange rate, would have to vary in an optimal way. In this case inferences would also have to be drawn regarding the coefficients of the model.

III. Financial Policies in a Two-Country World Economy

The discussion above suggests that in general circumstances it will always be

[7]Russell Boyer calculates an optimal decision rule in a model which is similar to the one employed here except that he assumes that home and foreign securities are perfect substitutes Frenkel's optimal rule is derived in a quite different model.

optimal for an individual country to opt for a managed floating exchange rate rather than a fixed or freely floating exchange rate if the authorities in the other country in a two-country world economy set their output at its full-employment level while pegging their interest rate. This result continues to hold when some types of modifications are made in the treatment of the foreign country. It could be assumed that foreign output and the foreign interest rate are random variables which may or may not be correlated with each other and with the shocks to the market relations considered so far (see Stephen Turnovsky). Alternatively the system could be expanded to comprehend two countries with the addition of appropriate market relations and additive disturbances, and it could be assumed that the home authorities know all the coefficients of the model, the joint distribution of all the additive disturbances, and the unchanging values of the policy instruments of the foreign authorities (see Robert Flood and C. Michael Jones).

Another approach to the analysis of financial policy in a two-country world is to determine whether or not two countries, each of which is committed to a particular monetary policy could agree on a pure exchange rate regime. Consider a two-country generalization of the system presented above with enough additive disturbance terms to permit analysis of shifts between every pair of the four financial markets, shifts between the markets for the two goods, and shifts in the market for each good alone.[8] Suppose that the authorities in both countries fix their interest rates. Both countries will prefer a fixed exchange rate if there are disturbances only in financial markets. Under plausible assumptions both countries will prefer a freely floating rate if there are shifts only between the markets for the two goods, since the country which undergoes the increase (decrease) in demand experiences an appreciation (a depreciation) of its currency which mitigates the effect of the disturbance on output. However, when shocks affect only the market for one good, the country producing it will prefer a freely floating exchange rate, while the other country will prefer a fixed exchange rate since movements in a freely floating exchange rate mitigate the output effects in the producing country but exaggerate them in the other country. As before, if all types of shocks must be faced, then the expected losses for each country associated with each of the two exchange rate regimes could be calculated and compared.

Suppose the authorities in each of the two countries can observe and react to all of the financial variables not chosen as policy instruments. Taken together the two sets of authorities can choose as policy instruments any three of the following six financial variables: the two money supplies, the difference between their holdings of foreign exchange reserves measured in the same currency, the two interest rates, and the exchange rate. If the authorities are concerned only about output deviations and if the coefficients of the system are known to them, then they can set one policy instrument arbitrarily. They should choose two linear decisions rules which tell them how to set the other two policy instruments given the levels of the three financial information variables. There is no conflict of interest between the two sets of authorities, and financial policymaking can be cooperative or decentralized as long as there is agreement about which policy instrument to keep fixed and how to do it.

The situation is quite different if each country has two objectives, say minimizing squared deviations in output and interest sensitive consumption from desired values, and if these values are inconsistent (as they will be in general when there are only three independent policy instruments). If the two sets of authorities behave like Cournot duopolists, no equilibrium exists. If one set of authorities or the other is allowed to behave like a Stackelberg leader, an equilibrium is reached which in general is off the contract curve (see Koichi Hamada, Jones, and Jürg

[8] In a beginning-of-period-balancing model, a shift which affects only the home good market could result from a reduction in home saving which is spent entirely on the home good; under plausible assumptions the qualitative effects would be the same in the case in which spending on both the home and foreign goods increases. Richard Sweeney uses an end-of-period-balancing model in which a change in saving behavior must be matched by a change in at least one asset demand function.

Niehans). In these circumstances both countries might be willing to agree either to refrain from changing their foreign exchange reserves or to use them in a well-defined way to fix the exchange rate since such an agreement might lead to better outcomes than those which would emerge under unrestrained noncooperative behavior. Though it is not obvious how to go about constructing them, it is possible that other simple guidelines for the management of foreign exchange reserves or the exchange rate would generate outcomes superior to both unrestrained noncooperative behavior and either pure intervention policy. Even if the two countries agree to a pure intervention policy, in general a policy conflict remains, and both countries can be made better off by cooperation.

IV. A Concluding Reminder

This discussion of financial policies has proceeded under simple assumptions about how private agents form their expectations; exploration of the implications of more sophisticated assumptions is important (see Flood and Michael Parkin). It has been assumed that there are no costs associated with changing the values of policy instruments. The short-run focus has precluded consideration of how the financial authorities should respond to the dynamic effects of saving, capital accumulation, the transfer of wealth between countries through current account imbalances, and monetary policies implying differing secular rates of inflation. Attention has been devoted to the financial policy problems which are evident when there are financial relations between only two countries; additional "optimum currency area" problems arise when there are relations among many countries (see Tower and Willet).

REFERENCES

R. S. Boyer, "Optimal Foreign Exchange Market Intervention," *J. Polit. Econ.*, forthcoming.

W. C. Brainard, "Uncertainty and the Effectiveness of Policy," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 411–25.

Ralph C. Bryant, *Money and Monetary Policy in an Open Economy*, Washington, forthcoming.

S. Fischer, "Stability and Exchange Rate Systems in a Monetarist Model of the Balance of Payments," in Robert Z. Aliber, ed., *The Political Economy of Monetary Reform*, London 1977.

R. P. Flood, "Capital Mobility and the Choice of Exchange Rate Systems," *Int. Econ Rev.*, forthcoming.

J. A. Frenkel, "International Reserves Under Alternative Exchange Rate Regimes and Aspects of the Economics of Managed Float," *Kredit und Kapital*, forthcoming

B. M. Friedman, "Targets, Instruments, and Indicators of Monetary Policy," *J. Monet Econ.*, Oct. 1975, *1*, 443–74.

K. Hamada, "Alternative Exchange Rate Systems and the Interdependence of Monetary Policies," in Robert Z. Aliber, ed., *National Monetary Policies and the International Financial System*, Chicago 1974

D. W. Henderson, "Modelling the Interdependence of National Money and Capital Markets," *Amer. Econ. Rev. Proc.*, Feb 1977, *67*, 190–99.

C. M. Jones, "Policymaking Efficiency and the International Monetary System," unpublished doctoral dissertation, Yale Univ 1978.

J. Kareken, T. Muench, and N. Wallace, "Optimal Open Market Strategy: The Use of Information Variables," *Amer. Econ. Rev.*, Mar. 1973, *63*, 156–72.

Robert A. Mundell, *International Economics*, New York 1968.

J. Niehans, "Monetary and Fiscal Policies in Open Economies Under Fixed Exchange Rates: An Optimizing Approach," *J. Polit Econ.*, July/Aug. 1968, *76*, 893–920.

M. Parkin, "A Comparison of Alternative Techniques of Monetary Control Under Rational Expectations," unpublished paper, Univ. Western Ontario 1978.

W. Poole, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quart. J. Econ.*, May 1970, *84*, 197–216.

J. L. Stein, "The Optimum Foreign Exchange Market," *Amer. Econ. Rev.*, June 1963, *53*, 384–402.

R. J. Sweeney, "Automatic Stabilization Policy and Exchange Rate Regimes: A General Equilibrium Approach," unpublished paper, U.S. Treasury 1976.

E. Tower and T. D. Willet, *The Theory of Optimum Currency Areas and Exchange Rate Flexibility*, Special Papers in International Economics, No. 11, Princeton Univ. 1976.

S. J. Turnovsky, "The Relative Stability of Alternative Exchange Rate Systems in the Presence of Random Disturbances," *J. Money, Credit, and Banking*, Feb. 1976, *8*, 29–50.

The Current State of the Policy-Ineffectiveness Debate

By BENNETT T. MCCALLUM*

The debate in question is, of course, over the applicability to the *U.S.* economy of the famous and controversial "neutrality" proposition—due primarily to Robert Lucas, Thomas Sargent, and Neil Wallace—according to which the choice among monetary policy feedback rules is irrelevant for the stochastic behavior of the unemployment rate in a neoclassical economy with rational expectations. Since the basic logic of this proposition has become well-known, I will not devote space to a formal statement or proof. It will be necessary, however, to inject some interpretive comments. This need arises because formal proofs of the proposition refer to effects of alternative policy rules on the stochastic *steady-state* behavior of key macro-economic variables. Thus the alternative policy rules are treated as permanently maintained, with transitional effects ignored. Under this interpretation the proposition does not imply that *actions* of the Fed (the "monetary authority") have no impact on unemployment rates. Since expectations at any moment of time are given, the more expansionary are the Fed's actions "this month" the lower will be "this month's" unemployment rate, even if the proposition is valid. Evidently, what the latter does suggest is that the Fed's choice among alternative reaction patterns, sustained over many periods of time, will have negligible effects on the level and variability of unemployment rates averaged over these periods. The distinction between policy actions and policy rules is crucial.

The proposition is straightforwardly applicable, therefore, only to hypothetical, maintained situations. One might even say that the proposition is a "long-run" result, though such terminology hardly seems helpful. But it does not follow that the proposition is irrelevant for actual policy, as some commentators have suggested. Clearly, it bears upon the following question: Does it matter, on average, if the money stock is typically increased rapidly, slowly, or not at all when high unemployment rates are observed? The proposition may apply only to the systematic, nontransitory component of policy behavior, but presumably that is the portion of primary interest to economic scientists and public-spirited policymakers.[1]

A few more matters should be clarified before I begin with the main discussion. In particular, it should be mentioned that the neutrality proposition refers to real output rates, as well as unemployment rates, but with the relevant concept in both cases measured relative to some "capacity," or "full employment," or "natural rate" benchmark. There are many models in which the proposition is valid for such measures even though capacity levels are themselves affected by the choice among policy rules. By taking these relative concepts as the ones under discussion, I will be focusing on "stabilization policy" and abstracting from issues of "economic growth." Notationally, the same symbol, y_t, will be used to refer to both concepts—

*Professor of economics, University of Virginia. I am indebted to the National Science Foundation for financial support under grant no. SOC 76-81422.

[1]That most "policymakers" are not, in their day-to-day activities, concerned with the choice among sustained feedback rules hardly diminishes the importance of the proposition under discussion. With respect to economists, my position has received implicit support from the practice of leading activist critics of the proposition: the analyses of Stanley Fischer, Edmund Phelps and John Taylor, and Taylor all focus upon stochastic steady-state behavior. (These examples also illustrate that the desirability of conducting policy by means of feedback rules is not at issue.) This sort of analysis abstracts from effects of two kinds: those due to initial conditions and those that occur while agents are in the process of learning about newly adopted policy rules. With respect to the latter, it is important to keep in mind that it is doubtful whether output or employment effects obtained by policy deception would be welfare enhancing

unemployment and (the *log* of) output relative to natural-rate levels. Finally, it will be presumed that information on period t values of all relevant variables becomes available in period t + 1, both to the Fed and to individual agents.[2] Thus the theoretical discussion presumes that the Fed has no informational advantage. It is generally agreed that such an advantage would invalidate the neutrality proposition without providing a strong basis for policy activism. See Robert Barro (1976).

I. Objections to the Neutrality Proposition

Let us now proceed by considering some of the main objections that have been raised by activist critics to the application of the proposition to the *U.S.* economy. For a while there was resistance to the notion that expectations are formed rationally, especially in the sense that all agents act as if they knew the true structure of the economy including the policy feedback rule.[3] More recently, however, macro-economic researchers seem to have moved toward a sort of implicit agreement that this extreme rational expectations assumption is appropriate for analysis of stabilization policy. There are, I believe, two main justifications for this view. First, there is no reason to believe that the assumption is terribly inaccurate, empirically, at the macroeconomic level.[4] Of course it is literally untrue, but so is every behavioral relation in every formal economic model. Second, every alternative assumption has an extremely unattractive property: it requires the assumed existence of some particular pattern of *systematic* expectational error. One would not expect any systematic pattern of errors to persist, however, since each would imply the existence of unexploited opportunities for enormous entrepreneurial gain. The relevant issue is not whether expectations are "actually" formed rationally, but whether it would be fruitful to conduct stabilization analysis under any other assumption.

One of the most prominent activist arguments for the inapplicability of the proposition has to do with "persistence" of positive or negative values of y_t (output or unemployment). According to this viewpoint, the proposition implies that y_t values are serially uncorrelated, so the fact that measured *U.S.* unemployment and output values exhibit strong serial dependence immediately provides an empirical refutation. But, as several writers have noted, while the rationality assumption implies that expectational errors are serially uncorrelated, there is no such implication regarding values of y_t. To emphasize this, Sargent (1979) has developed an agent-maximizing, market-clearing model in which adjustment costs lead to an aggregate supply function of the form

$$(1)\quad y_t = a_o(p_t - E_{t-1}p_t) + \sum_{i=1}^{n} b_i y_{t-i} + \epsilon_t$$

where p_t is the *log* of an aggregate price index, $E_{t-1}p_t$ is its conditional expectation, and ϵ_t is a disturbance uncorrelated with past values of all variables. In the example explicitly worked out, Sargent takes $n = 1$, but the analysis could in principle be extended to the more general case. Now (1) is clearly a supply function consistent with standard proofs of the proposition. And for many parameter values (1) will imply positive serial correlation of y_t values.[5] Furthermore, Lucas (1975b) has developed a model, quite different from Sargent's, in which neutrality prevails yet persistence results from certain

[2]For simplicity, the discussion in this paper will be restricted to aggregate relationships. Aggregate supply functions like equation (1) below are typically rationalized by means of the analysis of Robert Lucas (1973), in which agents in separate markets carry out transactions knowing values of aggregate variables for previous periods but only "local" absolute prices for the current period

[3]Actually, the proposition permits the somewhat weaker assumption that expectations differ from the fully rational values by a random term uncorrelated with available data.

[4]Arguments based on expectational differences across individual consumers or firms amount to objections to macroeconomics, not rational expectations Any readers who are *fundamentally* unsympathetic to the rationality assumption are urged to consider the position expressed in Lucas (1975a).

[5]Sargent (1979) also shows that the neutrality proposition is, in a two-shift model, consistent with the apparent failure of real wages to move countercyclically.

restrictions on the information available to individual agents. And very recently Alan Blinder and Fischer have shown that persistence can be generated, without violation of the proposition, by inventory behavior of a plausible type. Thus the persistence objection is not well founded.

The other main line of argument on the activist side has to do with "sticky prices." One version begins with the observation that many models yielding the neutrality result—most notably, the model of Sargent and Neil Wallace—assume that prices are "perfectly flexible" in the sense that the price level adjusts in each period so as to equate aggregate supply and demand. But actual prices are sticky—the price level seems to adjust very slowly to eliminate conditions of excess supply or demand. Accordingly, the argument goes, the neutrality proposition is inapplicable to the *U.S.* economy. There are, however, two flaws with this argument. First, it is not entirely clear that the Sargent-Wallace model rules out price level stickiness of the type that has been documented: the model permits a many-period, distributed-lag response of the price level to changes in the money stock. Secondly, it is possible to construct models in which prices are sticky in a stronger sense and yet the neutrality proposition prevails. For these reasons, discussed more fully in my 1978 paper, the proposition cannot be disposed of by simply noting that actual prices move sluggishly.[6]

There is, however, a more persuasive version of the sticky-price argument, one that involves the notion of "long-term contracting." In particular, it has been shown, most notably by Fischer, that scope for activist monetary policy will exist—the neutrality proposition will fail—in an economy in which typical labor contracts fix nominal wages for two or more periods in advance. In such an economy, the choice of a policy rule will not affect the mean value of y_t—so unemployment cannot be kept permanently low—but will influence the variability of y_t. This is possible, even if monetary policy is fully anticipated, because the policy rule can take account of shocks that occur after a contract has set the wage for a (fixed) portion of the workforce. A similar result obtains, furthermore, for multiperiod price-setting arrangements (see Taylor) and such arrangements may be formal or implicit.

In my opinion, this line of argument constitutes the most telling objection to the neutrality proposition that has been advanced to date. Nevertheless, it is not entirely compelling. As Barro (1977b) has emphasized, the procedure by which employment is determined in Fischer's contracting scheme is Pareto suboptimal: other contracts could conceivably be written that would improve the welfare of both firms and households. Thus there is no solid economic rationale for the presumption that Fischer-type contracts are written.[7] And it seems unlikely that any such contracts would remain in force if the policy authorities were to try to exploit them to a great extent. Fischer recognizes both of these points but argues that contracts of the form used in his model are, in fact, of the type that exist in actual economies.

II. Formal Empirical Evidence

Given the considerations discussed in the previous section, it appears unlikely that the ineffectiveness debate can be resolved by means of purely theoretical arguments or casual empiricism. Recourse to formal econometric evidence would seem to be necessary.

Unfortunately, however, it has become apparent that it is extremely difficult to bring such evidence to bear on the issue in an effective way. In part, this is because the neutrality proposition is compatible with supply functions more general than (1)—in particular, with functions that include *lagged*

[6]A related argument invokes the concept of "disequilibrium." For a discussion of disequilibrium models, see the papers elsewhere in this issue on "Macroeconomics: An Appraisal of the Non-Market-Clearing Paradigm."

[7]Indeed, Barro's analysis leads him to suggest that "sticky wages, layoffs versus quits, and the failure of real wages to move countercyclically" may be merely "a facade with respect to employment fluctuations" (1977b, p. 316).

one-period expectational errors (lagged "innovations"), such as

$$(2)\quad y_t = \sum_{i=0}^{k} a_i(p_{t-i} - E_{t-i-1}(p_{t-i})) + \sum_{j=1}^{n} b_j y_{t-j} + \epsilon_t$$

The proposition is not, on the other hand, consistent with a formulation that includes multiperiod expectational errors, such as

$$(3)\quad y_t = \sum_{i=0}^{k} a_i(p_t - E_{t-i-1}(p_t)) + \sum_{j=1}^{n} b_j y_{t-j} + \epsilon_t$$

Simple inspection suggests that it could be difficult to distinguish empirically between formulations like (2) and (3).

Indeed, Sargent (1976b) has shown that, without the use of additional a priori information of *some* type, it is logically impossible to distinguish between (2) and (3) or, more generally, between models in which the neutrality proposition is valid and invalid. In other words, the neutrality property alone places no restrictions on time-series data taken from a single policy regime. This does not imply that attempts to test the proposition are hopeless, but it does emphasize the importance of giving careful consideration to the type of a priori information used in any test attempt.

At present, the two most well-known empirical studies are those of Sargent (1976a) and Barro (1977a). In the first of these, identifying restrictions are imposed by the adoption of a supply function of type (1): since (2) is more general, this amounts to the assumption that the parameters $a_1, a_2, \ldots, a_k$ are all equal to zero. (Actually, Sargent emphasizes innovations in policy variables, such as the *log* of the money stock, m_t, rather than p_t. Also, he notes that it may be appropriate to interpret y_t and m_t as *vectors* of variables.) While this specialization is not implied by the proposition, it seems plausible—why should past errors affect today's supply decision? In any event, with the additional assumption that the equation's disturbance term is free of serial correlation, this exclusion of lagged expectational errors permits Sargent to obtain the implication that y_t is, given the effect of lagged y's, uncorrelated with past values of policy variables. Accordingly, evidence consistent with this implication would seem to provide genuine empirical support for the neutrality proposition, even though inconsistent evidence could result either from the proposition's invalidity *or* from the presence of lagged innovations in the aggregate supply function. As it happens, Sargent's results are mixed, but reasonably consistent with the implication. It has been noted, however, that Sargent's auxiliary assumption regarding serial correlation is crucial to this procedure (see Christopher Sims and Robert Shiller) and Sims has argued that the assumption is arbitrary. So, while Sargent's results may not be totally uninformative, they should probably be regarded as shedding little light on the validity of the proposition.

Barro's (1977a) test approach is quite different. It involves explicit estimation of an equation analogous to (2) but with monetary innovations appearing in place of expectational errors in p_t and with additional "real" labor market variables in place of lagged values of y_t. Testable implications are obtained under the neutrality hypothesis by the exclusion from the unemployment rate (y_t) equation of certain exogenous variables that appear significantly in the equation used to explain monetary policy. This exclusion permits Barro to distinguish between the effects of monetary innovations and anticipated values of the money stock (Barro, 1977a, pp. 109–10). His results for 1946–73 (annual data) are strikingly favorable to the neutrality proposition: current and lagged monetary innovations are highly significant while anticipated components of m_t provide no incremental explanatory power.

On the surface, it might appear that Barro's results are open to criticism on the grounds of implausibility: strong effects on y_t are found for expectational errors made one and two years earlier. The estimated equation can be interpreted, however, as resulting from the elimination of other real endogenous vari-

ables from an aggregate supply function in which only the current monetary innovation is present (see the author, 1979). Thus this apparent objection is not telling. A more reasonable source of uneasiness over Barro's results is, I believe, their reliance on an accurate decomposition of money growth rates into anticipated and unanticipated components; analysis of policy behavior rules is not something with which macroeconomists have a great deal of experience. Also open to skepticism is Barro's explanatory variable (in the y_t equation) designed to reflect effects of the military draft. Still, Barro's results—recently augmented (1978) by evidence on price level behavior—are quite impressive.

A third empirical approach, as yet less well-known, has been suggested by Sargent (1976b) and implemented by Salih Neftci and Sargent. If the monetary authority's policy feedback rule *changes* at some point of time, there should be a shift in the distributed-lag relationship of y_t on actual m_t values if the neutrality proposition is true but not if it is false, with this implication reversed for a relationship between y_t and innovations in m_t. Using quarterly *U.S.* data for 1949–74, Neftci and Sargent located a policy break at the start of 1964 and then obtained Chow-type test statistics that are reasonably supportive of the neutrality proposition. The main weaknesses of this procedure are the absence of any formal statistical basis for reaching conclusions and (again) the difficulty of characterizing policy behavior. The approach would seem, nevertheless, to warrant additional attention.

III. Conclusions

For the most part, the formal econometric evidence developed to date is not inconsistent with the neutrality proposition. But the power of existing tests is probably not high and, in any event, the evidence is not entirely clear-cut. Thus many economists may tend, at least for the present, to maintain adherence to their favorite theoretical model—whichever one offers the combination of features that seems essential. There is room for hope that future research will offer new insights, but it is hard to imagine that any conclusive breakthrough will occur. Thus it may be best to conclude by noting the extent to which the current brand of policy activism has been affected by the analysis and findings of the Lucas-Sargent-Barro school. Just over a decade ago, Milton Friedman's suggestion that unemployment could be kept low only by accelerating inflation seemed radical; now even many activists doubt that it can be kept low by *any* monetary policy stance.

REFERENCES

R. J. Barro, "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

———, (1977a) "Unanticipated Money Growth and Unemployment in the United States," *Amer. Econ. Rev.*, Mar. 1977, *67*, 101–15.

———, (1977b) "Long-term Contracting, Sticky Prices, and Monetary Policy," *J Monet. Econ.*, July 1977, *3*, 305–16.

———, "Unanticipated Money, Output, and the Price Level in the United States," *J Polit. Econ.*, Aug. 1978, *86*, 549–80.

A. S. Blinder and S. Fischer, "Inventories, Rational Expectations and the Business Cycle," M.I.T. work. paper no. 220, June 1978.

S. Fischer, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Feb. 1977, *85*, 191–205.

R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

———, "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ Rev.*, June 1973, *63*, 326–34.

———, (1975a) Review of *A Model of Macroeconomic Activity, Vol. 1*, by R. C Fair, *J. Econ. Lit.*, Sept. 1975, *8*, 889–90

———, (1975b) "An Equilibrium Model of the Business Cycle," *J. Polit. Econ.*, Dec 1975, *83*, 1113–14.

B. T. McCallum, "Price Level Adjustments and the Rational Expectations Approach to Macroeconomic Stabilization Policy," *J. Money, Credit, Banking*, Nov. 1978, *10*,

418–36.

——, "On the Observational Inequivalence of Classical and Keynesian Models," *J. Polit Econ.*, Apr. 1979, *87*, forthcoming.

S. **Neftci and T. J. Sargent,** "A Little Bit of Evidence on the Natural Rate Hypothesis from the U.S.," *J. Monet. Econ.*, Apr. 1978, *4*, 315–19.

E. S. **Phelps and J. B. Taylor,** "Stabilizing Powers of Monetary Policy under Rational Expectations," *J. Polit. Econ.*, Feb. 1977, *85*, 163–90.

Thomas **J. Sargent,** (1976a) "A Classical Macroeconometric Model for the United States," *J. Polit. Econ.*, Apr. 1976, *84*, 207–37.

——, (1976b) "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," *J. Polit. Econ.*, June 1976, *84*, 631–40.

——, *Macroeconomic Theory*, New York 1979.

—— **and N. Wallace,** " 'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

R. J. Shiller, "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review," *J. Monet. Econ.*, Jan. 1978, *4*, 1–44.

C. A. Sims, "Exogeneity and Causal Ordering in Macroeconomic Models," in *New Methods in Business Cycle Research: Proceedings from a Conference*, Minneapolis 1977.

J. B. Taylor, "Estimation and Control of a Macroeconomic Model with Rational Expectations," *Econometrica*, forthcoming.

A Case for Monetary Reform

By JAMES L. PIERCE*

Any paper on *U.S.* monetary reform must consider reform of the Federal Reserve System. This paper considers reforms of the Federal Reserve that should enhance the quality of monetary policy. Two kinds of reforms are considered: 1) changes in the internal institutional structure of the Federal Reserve that should enhance the quality of its monetary policy decisions; 2) changes in the powers of the Federal Reserve to impose reserve requirements that should enhance the efficacy of the policies themselves.

I. Internal Reorganization

On July 17, 1978, Senator William Proxmire released, during hearings of the Senate Banking Committee, the contents of a letter written to him in March by David M. Lilly, a former member of the Board of Governors. In that letter, Mr. Lilly described four specific areas in which he thought the Fed should be reorganized. The contents of the letter provide an excellent vehicle for discussing organizational reform of the Fed. In fact, to have the Lilly letter made public is an important reform itself; former governors of the Fed have been remarkably silent about flaws in that institution. Mr. Lilly's observations will be quoted, and then I shall offer my own comments on those observations.

1) *Organization of the Open Market Committee*

"I think that the presidents of the Federal Reserve banks should have the same accountability that applies to members of the Board as regards the Open Market Committee. The Reserve Bank presidents are neither appointed by the President of the United States nor by the Board of Governors,—yet they serve on the Open Market Committee and have input into monetary policy and on a rotating schedule vote on decisions that are crucial to the nation's well-being."

"Furthermore, with regard to monetary policy they are not accountable to the Board of Directors of their Reserve banks. Those Boards are excluded from monetary policy discussions connected with the OMC. Thus, in my view the Reserve Bank presidents are not responsible to anyone for their votes The accountability of the Reserve Bank presidents should be established if they are to continue to have a say in monetary policy."

The Federal Open Market Committee (*FOMC*), or OMC as Mr. Lilly calls it, is the primary vehicle for monetary policy in the United States. It makes all the decisions concerning the execution of open-market operations. These operations in turn are directed toward affecting the growth of money and credit, the level of interest rates and, ultimately, the level of economic activity in the United States. Changes in reserve requirements and in the discount rate, which are determined by the Federal Reserve Board, are distinctly secondary to open market operations in the conduct of *U.S.* monetary policy.

The *FOMC* has twelve voting members: the seven governors of the Federal Reserve, the president of the Federal Reserve Bank of New York, and four presidents of the remaining eleven Reserve Banks; these four presidents serve on a complex rotating basis. All twelve Reserve Bank presidents take part in *FOMC* meetings but only five vote on policy. Of the twelve votes on the *FOMC* only the seven from the Board of Governors are cast by individuals who receive presidential appointments and Senate confirmations. Reserve Bank presidents are not appointed by any public official. Rather, they are appointed by the private directors of their Reserve Banks with the appointment confirmed by the Federal Reserve Board. The Bank directors are private citizens; one third of whom are bankers, one third are individuals selected by

*University of California-Berkeley

bankers, and one third are selected by the Federal Reserve Board.

Mr. Lilly observes that Federal Reserve Bank presidents are accountable to their boards of directors, but these directors are not privy to monetary policy discussions or decisions. Thus, because the directors are not aware of what the bank president said in *FOMC* meetings and only learn of his vote after a significant time delay, the Reserve Bank presidents actually are accountable to no one. He argues that their accountability should be established. I agree, but changing the nature of appointment is not sufficient to establish accountability.

Many proposals have been made over the years to increase the accountability of the Fed and the *FOMC*. There are two elements in this accountability: who appoints the decision makers and to whom must they explain their actions once appointed. Mr. Lilly only addresses the first element. It will be discussed briefly here before turning to the second element.

The question of who appoints the members of the *FOMC* is the less important of the two elements. Nevertheless, most observers would agree that the private directors of Reserve Banks have no business appointing members of the *FOMC*. Many proposals have been made over the years to rectify the situation. Some would restrict the *FOMC* to Federal Reserve Board members, others would retain the participation of Bank presidents on the *FOMC* but require either presidential or Federal Reserve Board appointment of the presidents with Senate confirmation in either case. Since I prefer restricting the *FOMC* to Federal Reserve Board members, I should like to point out some of the difficulties involved in having Reserve Bank presidents sit on the *FOMC*.

The Federal Reserve Board currently can exert a powerful influence over selection of Reserve Bank presidents by confirming or rejecting names supplied to it by the bank directors. Further, and perhaps more important, the Federal Reserve Board approves the budgets of the Reserve Banks. The Board and/or its chairman can make life very unpleasant for a Bank president who causes too much trouble at *FOMC* meetings. Thus, even if Bank presidents were appointed by the president of the United States and confirmed by the Senate, they would not be free agents so long as the Board controls their budgets. There are Bank presidents who act independently of the Board, but by-and-large, Reserve Bank presidents go along with the Board.

It is difficult to overstate the power that the Board of Governors, and particularly its chairman, has in the *FOMC*. Perhaps a few examples will make the point. The Board controls (or at least it did under the previous chairman) all staff material presented at the *FOMC*, and only Board staff members make verbal presentations at *FOMC* meetings. Reserve Bank presidents usually seek Board approval prior to agreeing to testify before Congress. Reserve Bank presidents typically have their testimony reviewed by Board staff and the Board itself prior to delivery. Board members do not check their testimony with Bank presidents prior to delivery. Finally, Bank presidents do not form coalitions within the *FOMC* but the Board often represents such a coalition. It is unlikely that a presidential appointment would make Reserve Bank presidents much more willing to buck the Board and its chairman.

There appear to be two solutions to the nonindependence of Reserve Bank presidents. The first would simply accept the reality of the situation and exclude Reserve Bank presidents from the *FOMC* and monetary policy decisions. This solution would have the virtue of fixing responsibility directly with the Board of Governors. It would have the deficiency of depriving monetary policy deliberations of regional influences and knowledge. The second solution would retain the Bank presidents on the *FOMC*, require presidential appointments with Senate confirmation *and* take the budget control over the Banks away from the Federal Reserve Board. About the only feasible way that budget control can be taken from the Board is to change the entire budgetary treatment of the Federal Reserve System, including the Board. This could be accomplished by placing all Fed expenditures within the federal budget. Such a change

would entail congressional authorization for Federal Reserve expenditures, both of the Board and the twelve district Banks.

Reform of appointment of members of the *FOMC* would enhance the accountability of that body. The most important improvement in accountability, however, must come from greater disclosure of decisions by the *FOMC*. Disclosure was not mentioned in Mr. Lilly's letter, but it is the key to accountability. The Fed cannot be held accountable until it is forced to announce what its policies are, how it selected them rather than others, what effects it expects to have with the policies and how it will modify them if events do not materialize as expected. In accountability, it is crucial to distinguish objectives and the methods selected to achieve them from honest policy mistakes that result from an uncertain environment. Disclosure is the only method of making this distinction. Some progress has been made to increase disclosure, first in the form of House Concurrent Resolution 133 and later in an amendment to the Federal Reserve Act which made most of the elements of the Resolution permanent. The Fed does announce its money growth targets, but it refuses to name objectives for the economy and to divulge its forecasts and policy alternatives. True accountability will occur only when these factors are disclosed.

2) *Board of Directors of the Reserve Banks*

"Only the three Class C directors are chosen by the Board of Governors. The Class A and B directors are chosen by the member banks. This ostensibly gives the member banks a larger voice in the running of the Reserve banks than the Board of Governors. In light of the reforms made with regard to the interests to be represented by members of Board of Directors made by Public Law 95-188, I believe it would be desirable to have both Class B and Class C directors selected by the Board of Governors in Washington."

Currently, two-thirds of the directors of Reserve Banks are chosen by member banks in the district and one-third are selected by the Federal Reserve Board. Recent changes in the law (P.L. 95-188) impose antidiscrimination standards on selection of directors and liberalize slightly the standards for selection. Despite these minor changes, the majority of Reserve Bank directors is still selected by member banks in the district. Mr. Lilly would like to have all directors chosen by the Board of Governors in Washington.

It is easy to make too much of the issue of Reserve Bank directors. They really exert no influence on monetary policy; their primary function is to advise the Bank president on internal operations and to provide community involvement with the Federal Reserve. Bank directors might appear to have a policy role because requests for changes in the discount rate must come from Reserve Banks, but the Board of Governors is free to ignore these requests and usually does. The discount rate function does lead to economic briefings of the board of directors so that possible submissions of discount rate changes can be considered. Because the Board in Washington must approve these requests and because most requests are not even considered, the role of the bank directors is not important. The Board almost always has a menu of previously proposed changes in the discount rate available, should it wish to change the rate. If there is nothing on the menu that the Board likes, a phone call solves the problem.

Central banking is a governmental function and it is inappropriate to have private individuals serving as directors of Reserve Banks. Central banking functions are no more important to bankers than to anyone else, yet the selection of directors is dominated by bankers. Mr. Lilly would like to have the bank directors selected by the Board in Washington; I would like to see them eliminated altogether. If a Reserve Bank president wants help with internal operations, then let him appoint an advisory group.

3) *Deferral of Open Market Committee Directive*

"I see no reason why the release of the policy directive of the OMC needs to be delayed. Everyone should have the same access to the decisions made by the OMC. Currently, only those brokers and dealers with large staffs monitoring Federal Reserve policy on a daily, and in some cases hourly,

basis can know what monetary policies the OMC is pursuing. This is discriminatory and gives brokers and dealers an advantage over the ordinary citizen."

The deferral of release of the *FOMC* directive has been a hotly debated issue for some time. Some people, many of whom are in the Fed, argue that speedy release of the directive would be harmful because it would encourage speculation. It is difficult to find merit in this argument. Insiders such as government security dealers know very quickly when the Fed has changed policy. After all, the Fed executes policy through open market operations with these dealers. Other large operators in the money market employ Fed watchers who have become very good at divining when the Fed has changed policy. It is difficult to understand why the rest of the public must wait thirty days to learn of Fed policy.

There appears to be a belief within the Federal Reserve that secrecy and confusion about current policy enhances the effectiveness of that policy. I know of no basis for this belief. The sooner the public knows what monetary policy is, the better. The public cannot decide what to do with information until it has it.

I believe that the real reason the Fed defers release of the directive is its penchant for secrecy, which in turn is a desire to avoid accountability. If the Fed truly had its way, I doubt that it would ever release the directive, it would usually produce only platitudinous statements about the "thrust" of policy.

Speedy release of the directive is clearly called for. While I think it is helpful to have policy debates in private in order to invite free interchange of ideas, once decisions are made they should be announced immediately.

4) *Monetary Policy Responsibilities of the Board of Governors*

"I accepted the position on the Board because I viewed, and still do, the Board's monetary policy responsibility to be of utmost importance to the nation. Unfortunately, there are many other matters that come before the Board that are time consuming, and these detract from this major responsibility."

Contrary to popular opinion, the Federal Reserve Board and Reserve Bank presidents spend most of their time on matters other than monetary policy. The Federal Reserve Board spends most of its time on bank and holding company regulation. The Reserve Bank presidents are concerned not only with regulation but also check clearing, funds transfer, and other operating activities. It seems reasonable to assert that monetary policy is a full-time job and policymakers should not be distracted by other matters.

The Federal Reserve is on both sides of this issue. On the one hand, many Board members have felt the frustration indicated by Mr. Lilly over the relatively small amounts of time available for monetary policy. On the other hand, the Board has resisted efforts to reduce its regulatory burdens. When faced with the prospect of seeing its regulatory functions go to other agencies, the Fed evidences a strong desire to protect its turf.

The Fed has argued strenuously that it needs regulatory functions in order to help it execute monetary policy. I know of no case in which monetary policy was helped by having the Fed in the regulatory business. I know of many cases in which regulatory responsibilities got in the way of monetary policy. The Federal Reserve needs data on what is happening with respect to banks and in financial markets in general. It does not need to regulate in order to obtain these data. I believe monetary policy would be significantly improved if the Fed ceased being a regulator.

II. Some Further Considerations

Mr. Lilly's complaints seem well founded. They all spring from the same institutional source. The basic problem lies with viewing the Federal Reserve as a banking agency. A central bank is not a private bank; it plays its role by affecting the nation's monetary base.

The current structure of the Fed has its roots in history, not in good economics. It was history that produced Reserve Banks that were set up like private banks with stockholders (member banks) and boards of directors.

The stock of the Reserve Banks should be retired (purchased from member banks) and the boards of directors eliminated. Reserve Banks should become purely governmental entities.

It was also history that made membership and regulations by the Fed come with reserve requirements. It is important to divorce reserve requirements from Fed membership and regulation. Required reserves held at the Fed are helpful to monetary policy. Reserve requirements should have nothing to do with the type of charter an institution has or with who regulates it. If reserve requirements should be imposed on a particular kind of liability for purposes of monetary policy, then they should be imposed on any institution that accepts that liability: member bank, nonmember bank, savings and loan, mutual savings bank, or credit union.

The Federal Reserve has found itself with declining membership primarily because many banks have found required reserves onerous. It has used all sorts of schemes to attract members. The Fed doesn't need members, it needs authority to impose reserve requirements.

While there is no evidence that declining membership has injured monetary policy, there is evidence that the Fed's Rube Goldberg graduated reserve scheme has harmed monetary control. There is also reason to believe that with automatic transfer accounts starting in November, and with nationwide *NOW* accounts (or their equivalent) waiting in the wings, there could be explosive growth in transactions accounts offered by institutions that do not have reserve requirements imposed against them. If this explosion occurs, the Fed could find its monetary policy control slipping. The solution appears to lie with imposing reserve requirements against all transactions accounts and allowing all institutions that offer them to have full access to Fed services including the discount window. The answer does not lie with forcing these institutions to be members of the Fed. If all institutions have access to Fed services, there will be no incentive to be members and the Fed's regulatory burdens should die a natural death. This reform would solve many of Mr. Lilly's problems.

The Economics of the Oceans: Environment, Issues, and Economic Analysis

By MAURICE WILKINSON*

The purpose of this paper is to review briefly the changing ocean environment, ocean policy issues, and some of the areas of economic theory and measurement most relevant to ocean economics.

I. The Environment

From the seventeenth century until the end of the nineteenth century, ocean politics and law reserved the oceans as open space. The necessary conditions for the maintenance of this open system were the military and economic domination of the world by the powers who were best served by open access, and the relatively slowly evolving technology of sailing ship construction and fishing methods which underlay the techniques of use and exploitation of the oceans.

Signs of movement away from the open system can be detected as early as the nineteenth century with, for example, the early recognition of the need to conserve fish stocks (see Giulio Pontecorvo and the author). By the 1930's, improvements in the technology of catching fish created the first long-distance fishing fleets. The dramatic shift away from resource abundance to resource scarcity, however, has come since the end of World War II. Under the pressure of increased population and national economic development throughout the world, and especially from the impact of the rapidly increasing capability to exploit ocean resources, the open ocean system is being transformed, substantially circumscribed, until it stands today in peril of elimination.

*Professor, Graduate School of Business, Columbia University. A longer version of this paper containing specific references to the literature is available from the author

At this point in time, the full extent and the form of national domination of ocean space and resources is not fully apparent. But it is fair to say that the question of international income distribution which has, since its inception, underlain the specific issues at the Law of the Sea Conference is in large part being resolved by the extension of coastal states' rights over ocean space. This situation represents an increase in the potential income of ocean states at the expense of states not so geographically fortunate. Thus far all participants in the sea law debates have denied the possibility of restrictions on trade and commerce, yet the extension of jurisdiction raises the possibility that in time restrictions may be imposed. For many states, the national interest is consistent with a set of complex regulations with respect to fisheries, minerals, pollution, and supervision and maintenance of straits.

II. Policy Issues

Almost all economic policy issues essentially involve the design of a regulatory mechanism to promote more rational management of ocean resources (see Pontecorvo, D. Johnston, and the author). Unfortunately, the history of economic regulation of the oceans has generally been characterized by incorrect management goals, improper regulations, inadequate information, and neglect of the research that must support proper regulation.

A. *Issues Concerning Extended Economic Zones*

At the present time the most serious policy issues involving the management of resources

in extended economic zones concern fisheries, coastal zone management, and petroleum. Conflict is highest with regard to fisheries, which currently have the highest economic value and the longest history of exploitation. The principal issues concerning management of fisheries within economic zones of developed countries continue to be the following: 1) the formulation and analysis of bioeconomic models of multispecie fisheries; 2) the characteristics of the supply functions for individual fisheries; 3) the cost and benefits of regulating fisheries; 4) the organization, structure, and political effectiveness of the fishing industry.

1) Bioeconomic control has as its objective the maximization of net revenue from utilization of the resources of the biomass. Focusing biological management on the biomass rather than individual fish stocks is a departure from most existing practices. Explicit recognition is given to the possibility that there may be a set of possible biomasses and the purpose of bioeconomic control is to define and implement criteria for choosing the particular biomass that will maximize social welfare. The common property characteristic of fisheries results in inefficient allocation of resources if exploitation is undertaken with a competitive market organization.

2) At the present time most species in *U.S.* and Canadian fisheries that have been commercially exploited are on the endangered species list compiled by the National Marine Fisheries Service (salmon is one exception). Individual stocks have been successively depleted as fisherman move from one stock exhibiting falling yield to the next most economically promising specie. Offshore fishing by long distance travelers from the *USSR*, Japan, etc. has also had a considerable impact upon all the fish stocks. Biologists are not able to accurately analyze the relative impact of domestic and foreign fishing on the stocks. Furthermore, it is not clear whether these stocks have been subject to general biological overfishing or an uneconomical level of exploitation. That is, the management criterion utilized during the last two decades has been "maximum sustainable yield." This principle does not result in an optimum economic result for domestic fishing industries, nor perhaps for foreign fishing. In short, there is considerable uncertainty concerning the long-run supply functions of individual species due to both a lack of biological knowledge and a lack of economic analysis in the management of fisheries.

3) A third key issue concerning fisheries is the benefits and costs of regulating multispecie fisheries. Given the lack of biological knowledge, the cost of acquiring additional empirical observations, the administrative expense of managing complex regulatory schemes (taxation and/or artificial markets), and the uncertainty concerning potential supply (and the generally poor prior performance of past efforts to manage fisheries), many fisheries might be better consigned to "benign neglect."

4) The present and future structure of the fishing industry (degree of integration, capital stock, marketing and distribution, etc.), its relationship to the growing state and federal regulatory bodies (regional councils), and its performance relative to foreign fishing fleets need examining as a basis for economic policy toward this industry. The traditional industrial organization studies (with the emphasis on institutional and empirical analysis) that provide the underpinnings for business and government relations in other areas of the economy are lacking for this industry. It is not an extreme question to ask whether a country such as the United States should employ foreign fishing fleets for harvesting and retain only domestic processing and distribution.

The principal remaining policy issues concerning extended economic zones are mineral exploitation (gas and petroleum) and coastal zone management (recreation and pollution). Offshore petroleum fields are a publicly held asset currently being explored by domestic oil companies. The regulatory mechanism adopted largely duplicates that used for exploration on public lands. Competitive bidding for extended leases is intended to capture for public use the excess profits or economic rent. The deficiencies of this arrangement include the small number of

private companies involved (particularly the lack of new entrants to the industry), the lack of analysis of the allocative implications of alternative bidding schemes, and the proprietory nature of information regarding petroleum fields.

Given these shortcomings there is little reason to expect the auction price to approximate the present value of any rents. This has led to the proposal that the regulatory agency have the authority to impose a profits tax to ensure that the public shares any "unanticipated profits." Actually implementing such profits taxes, however, would not be easy since much of the information on costs and revenues are not currently public. Corporations may therefore attempt to disguise profits by shifting costs from less profitable operations and by expanding into other areas of economic activity (including other energy sources). The possibility of such problems has suggested to some observers the need for a public authority to engage in exploration and production (similar to the Tennessee Valley Authority in electrical power).

B. *International Issues*

Economic policy issues concerning ocean space outside extended economic zones are deep sea mining (manganese nodules—manganese, copper, nickel, and cobalt—oil and gas), shipping (pollution from ships), fishing (migratory species such as whales and tuna), and military uses of the oceans. The first three issues are all a direct reflection of the increasing value of ocean resources that was discussed above in Section I. The increased political activity with regard to these issues parallels the rising market value of the resources involved.

The control of the profits and technology from deep sea mining, the characteristics of the supply functions for individual minerals, and the impact upon world prices (and hence the economies of several *LDC*s) are major issues at the Law of the Sea proceedings. The technology only exists in the hands of private concerns in a few countries (the United States and France). Consequently, many developed and less developed nations favor an international authority to control both the distribution of profits *and* the technology. Potential control over deep sea mining by private monopolies raises many of the same issues as offshore oil and gas exploration.

III. Economic Analysis

The two most relevant bodies of literature for the analysis of ocean problems are the theory of economic externalities or public commodities (see for example Robert Russell and the author, ch. 18) and the optimal control or management of natural resources (see for example L. G. Andersen, ch. 6). The implications of economic externalities or public commodities for ocean problems were first investigated in the context of fisheries management. The literature that followed provides the theoretical basis for the regulation of fisheries in order to prevent their overexploitation (see for example, Andersen, chs. 4–5; Pontecorvo, Johnston, and the author, pp. 66–77). The existing propositions concerning endowment of property rights, optimal taxation, and artificial markets are equally relevant for the regulation of ocean mining, pollution, and shipping, although these problems have yet to receive the attention devoted to fisheries.

Optimal control or management models for natural resources are an application of modern capital theory. Such models seek to identify the possible deficiencies of static analysis under conditions of certainty by introducing dynamics (intertemporal profit or net social return maximization) and uncertainty concerning technology and demand. The crucial question is how deficient are traditional static models for actual resource management? The principal conclusion derivable from static analysis of fisheries is to reduce fishing effort by taxing output (see for example, Andersen, ch. 5). Exactly how this is to be accomplished is left to actual management experience. Dynamic models under uncertainty confirm this general conclusion and in addition promise to specify exactly how taxes should be altered over time in response to changing biological and economic conditions. However, most existing bioeconomic

control models generally resemble Keynesian macro-economic growth models and exhibit the same deficiencies: they are long-run steady-state models that do not specify the rate of adjustment of the biomass; furthermore most models have never been estimated with actual data. Hence at this time they are not useful management tools.

What of future research? The combination of the complexity of the underlying biological processes together with the economics of common property resources results in extremely difficult optimal control problems. Analytical results will not be possible except for extremely simplified cases, and numerical solutions will be expensive and very sensitive to error in functional specification and assumed parameter values. For example, concave criterion functions give rise to steady-state policies while nonconcave functions (for example, increasing returns to scale) give rise to cyclical fishing policies. In this situation management will have to operate in the near future with simpler rules of thumb (such as reducing fishing effort in response to reductions in yield and revenues from harvesting) until additional research is completed with regard to both data collection and empirical econometrics.

Data on ocean economic activity are gathered by individual nations and compiled by *UN* agencies such as the Food and Agriculture Organization. No nation currently produces a consistent set of estimates of the income and product resulting from ocean economic activity. The United States has made one prior attempt to estimate "the value of its oceans" and the National Income Division of the Department of Commerce is currently attempting to create a subdivision, "The Ocean Sector," of the national accounts. This will be derived for the year 1972 from those industries where the contribution of the oceans' resources to *GNP* may be measured. Finally, the United Nations publishes some estimates of general ocean output.

Empirical studies of the demand and supply of ocean output are for all purposes limited to fishing and to a lesser extent shipping. The literature contains a few examples of demand studies that analyze separate food commodities (see for example Anderson and the author). There has been little estimation of supply and cost functions for individual fisheries. (Some exceptions are quoted in Andersen.) Empirical models for the management of individual fisheries are also few in number.

Finally, there are no econometric models of the fishing industry comparable to those that have been developed and utilized for policy analyses in related commodity markets such as agriculture (see Enrique Arzac and the author, 1979a, b, c).

The shortage of empirical research largely reflects the lack of biological knowledge and data. The consequences of this situation are that policy decisions are resolved in an arena devoid of the benefit (or constraint) of estimates of the economic benefits and costs of alternative courses of action.

Domestic *and* international political conflicts proliferate in such a setting where no competing party or interest group is really informed about the economic value of the resources involved. In the end the value of economics may be that it defines the economic dimensions of tradeoffs and compromises and thus encourages agreement. The growing conflict over ocean space and resources just might be reduced by this knowledge.

REFERENCES

L. G. **Andersen**, *The Economics of Fishing Management*, Baltimore 1977.

R. **Anderson and** M. **Wilkinson**, "An Evaluation of Alternative Consumer Demand Systems Within an Econometric Model of U.S Livestock and Feed Grain Markets," unpublished paper, Grad. Sch. Bus. Columbia Univ., Oct. 1978.

E. **Arzac and** M. **Wilkinson**, (1979a) "A Quarterly Econometric Model of U.S. Livestock and Feed Grain Markets and Some of Its Policy Implications," *Amer. J. Agr. Econ.*, 1979 forthcoming.

——— **and** ———, (1979b) "Stabilization Policies for United States Feed Grain and

Livestock Markets," *J. Econ. Dynamics Control*, 1979 forthcoming.

——— **and** ———, (1979c) "The Analysis and Control of Agricultural Commodity Markets," in A. Bensoussan et al., eds., *Applications of Stochastic Control Theory in Econometrics and Management Science*, Amsterdam 1979.

G. Pontecorvo, D. Johnston, and M. Wilkinson, "Conditions for Effective Fisheries Management in the Northwest Atlantic," in L. G. Andersen, ed., *Economic Impacts of Extended Fisheries Jurisdiction*, Ann Arbor 1977.

——— **and M. Wilkinson,** "From Cornucopia to Scarcity: The Current Status of Ocean Resource Use," *J. Ocean Develop. Int. Law*, forthcoming.

Robert Russell and Maurice Wilkinson, *Microeconomics: A Synthesis of Modern and Neoclassical Theory*, New York 1979.

The Economics of Marine Fisheries Management in the Era of Extended Jurisdiction: The Canadian Perspective

By PARZIVAL COPES*

In a sudden rush of unilateral declarations the 200-mile fishing limit in 1977 became the accepted norm in world practice and thus in international law. The Third Law of the Sea Conference—deadlocked on nonfisheries issues—so far has failed to produce a new international convention. But from its deliberations has emerged an *Informal Composite Negotiating Text* (*ICNT*), the fisheries provisions of which have been received with near-universal support or acquiescence. Several countries—including Canada and the United States—have indicated that they will adhere to the rules of the *ICNT* in administering their 200-mile fishing zones.

H. Scott Gordon's analysis that open access exploitation of a common property fish stock attracts excessive effort, leading to dissipation of resource rents, is now widely accepted. A major justification for extended jurisdiction is that it gives coastal states property rights over fish stocks that previously were international common property. By limiting access and managing effort wisely, coastal states may stop overexploitation and regenerate net economic benefits. Increasing pressure on the world's resources has greatly enhanced the potential value of the available fish stocks and thus the incentive for coastal states to claim the benefits of exploitation for their own nationals (see Francis Christy).

I. The New Regime

The principal effect of the *ICNT* with respect to fisheries is to accord to coastal states "sovereign rights for the purpose of exploring and exploiting, conserving and managing" the fish stocks in an "exclusive economic zone" extending 200 nautical miles outward from the coast. The "sovereign rights," however, are implicitly diminished by the requirement of a coastal state, where it "does not have the capacity to harvest the entire allowable catch," to "give other states access to the surplus." But as the coastal state is allowed unilaterally to determine the allowable catch (so as to leave a surplus that could be zero) and to set conditions of access (including fees that could be prohibitive), the coastal state is little constrained by formal rules. The real constraint lies in the need for a coastal state to be seen to be reasonable in its application of the *ICNT* rules, which are obviously intended to prevent the waste of fish that would result from underexploitation of available stocks.

The 200-mile limit is an imperfect instrument for rational resource management, insofar as it gives authority to states in relation to geographically defined areas rather than to distinct biological units. The 200-mile limit does encompass most of the world's continental shelves and slopes, so that the important demersal stocks that are tied to these shallower areas mostly come under coastal state management. Nevertheless, there are many important stocks that straddle or migrate across the outer limits of 200-mile zones or the border lines separating the adjacent zones of neighboring countries. Recognizing this contingency, the *ICNT* enjoins states concerned to cooperate in the joint management of transboundary stocks.

The *ICNT* makes special provision for "anadromous species" (mostly salmon), which spawn in the fresh water streams of "states of origin" and migrate beyond 200-mile limits to the high seas where they are potentially subject to a fishery by nonstates of

*Simon Fraser University. Research support from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged

origin (see the author, 1977). States of origin incur substantial (explicit and implicit) costs in maintaining salmon stocks. Their renewal is best managed by culling the schools as they gather in coastal waters to ascend their spawning rivers. Therefore, Canada and the United States, as states of origin for particularly valuable salmon fisheries, have been anxious to obtain recognition of their exclusive rights to anadromous stocks. The *ICNT* includes a provision recognizing that states of origin "have the primary interest in and responsibility for" anadromous stocks. The provision disallows high-seas fishing for anadromous species except where this "would result in economic dislocation for a state other than the state of origin." This presumably is meant to sanction the continuation of (limited) existing high-seas fisheries for salmon, but to proscribe any expansion.

While the enforcement authority of coastal states within 200-mile zones is widely recognized, it appears inconceivable that states should attempt to use police or military power to enforce fisheries claims beyond 200 miles. Any provision in international law—whether or not it is codified in a formal convention—applies only to states that submit to it. In practical terms, then, there is little a state of origin can do to stop high-seas fishing by other states for "its" salmon.

II. The Canadian Posture

In international fisheries relations Canada's position is that of a developed coastal state par excellence (see the author, 1972b). This country has the technical capability to exploit the extensive stocks available in its coastal zone and is economically motivated to reinforce the fishing economy of depressed coastal communities. With nearby resources far in excess of domestic needs and facing high opportunity costs for capital intensive operations, Canada is disinclined to engage in distant water fishing operations. Having also no great maritime power ambitions, Canada has shown an undivided concern for the protection of coastal fisheries and has taken a leading role in international deliberations promoting the interests of coastal states that are now expressed in the *ICNT*. In the process Canada moved by degrees to a fully acquisitive position on the 200-mile fishing limit. To do so required some modification to the country's traditional "internationalist" approach, with Pearsonian overtones of altruistic mission (see Allan Gotlieb and Charles Dalfen). Canada is now attempting to reconcile the pursuit of a distinctly national interest with a still enlightened international policy of cooperation and responsible relations.

The Canadian government in 1976 developed a comprehensive policy document that promised to apply economic rationality to the management of fish resources within the country's anticipated 200-mile zone. In putting the announced policies into effect, the government has been constrained by internal political considerations. The country's most extensive fisheries, on the Atlantic coast, have long been in a generally depressed state. The factors responsible include common property rent dissipation, escalation of foreign fishing effort leading to resource depletion, and greatly excessive labor inputs in the inshore sector related to the dearth of alternative employment opportunities (see the author, 1978). These conditions have resulted in a high level of subsidization of the fishing industry to support the local income base, as well as persistent demands to bar all foreign effort at the earliest opportunity.

Despite these pressures, the Canadian government has given fair consideration to distant-water fishing nations. Restrictions on foreign fishing operations have been phased in by stages and reasonable quota concessions have been made. Undoubtedly Canada's desire to retain diplomatic influence based on a favorable international image is part of the explanation. The country's trading interests are also well served by a reasonable stance on the 200-mile limit. But, most importantly, direct concessions in the fisheries area have been obtained in return for fishing quotas within the Canadian zone. These have included recognition of the abstention principle applied to salmon on the high seas, acknowledgement of Canadian primacy in setting management rules for stocks straddling or bordering the Canadian 200-mile

limit, concession of easy access for Canadian fish products to foreign markets and agreement to land some catches for processing in Canadian plants. In addition, Canada is now collecting substantial fees from foreign vessels fishing in its 200-mile zone.

III. Rational Management

From a global welfare-economic perspective the advantage of the 200-mile fishing limit lies in the circumstance that coastal states become sole owners of fish stocks within the limit, enabling them to introduce rational management. Specifically, by appropriately controlling effort, resource rents may be maximized. However, there is less reason to believe that the coastal states will be impelled to maximize other possible net benefits, such as consumers' and producers' surpluses, along with resource rents (see the author, 1972a). With fishing rights being concentrated in the hands of coastal states, the excess of catches relative to domestic needs is likely to increase, leading to a higher volume of fish products being destined for world trade. Exporting states are unlikely to take measures to constrain prices so as to enhance consumers' surpluses in importing countries. At the same time, coastal states are likely to neglect the producers' surpluses that may be lost in diverting fish resources from intramarginal operations in foreign fleets to marginal operations in domestic fleets. However, casual assessment leads me to conclude that the 200-mile limit is likely to cause much greater increases in resource rents than losses in consumers' and producer's surpluses.

Certainly, the 200-mile limit is helping to expand coastal state fishing operations at the expense of distant-water operations. In the long run this probably represents a superior allocation of resources. Intrinsically, there are great economic advantages to shore-based fishing over distant water fishing. The former permits more effective forms of specialization. Shore-based vessels may be designed for optimal catching efficiency without concern for bulky and expensive additional capacity to permit shipboard processing and lengthy storage. The latter functions may be carried out more effectively at a shore base with generous plant layout, adequate services, and a resident labor force requiring no extradomestic accommodation.

In Canada's case there are economically sound reasons for allocating some fishing quotas to distant water fleets, at least in the short run. These fleets represent sunk costs, as does the training of their crews and much of the infrastructure of their home bases. With the extension of fisheries jurisdictions around the world the alternative employment opportunities of distant water fleets are greatly shrunken. Until they are overcome by wear, tear, and obsolescence, distant water fleets will be operating with particularly low current opportunity costs. Moreover, the distant-water fleets have been engaged in some fisheries (for example, in deeper water and in ice conditions) for which Canada has not yet developed a fully adequate operational capacity. It makes good sense for Canada to allow foreign fleets access to stocks which its own fleets cannot harvest economically at this time—particularly in view of the valuable considerations, enumerated above, that Canada has been able to extract in exchange.

The scientific and administrative capabilities of the developed coastal states provide no assurance of a political will and economic adaptability to optimize fully the use of fish stocks under their control. The difficulties are well illustrated in the Canadian case. Prior to introduction of the 200-mile zone Canada already had some opportunities to rationalize east coast fishing operations by reducing effort applied to overexploited inshore stocks under national control (see the author, 1978) But under pressure of exceedingly high levels of unemployment in coastal communities, together with demands for financial support to bolster low fishing incomes, the government in fact engaged in an extensive program of subsidization. This maintained an excessive labor force and increased inshore fishing pressure, keeping net economic returns at a chronically depressed level, both in aggregate and per unit of effort terms.

The acquisition of property rights to all the stocks in Canada's east coast 200-mile zone now permits the diversion of excess fishing

effort inshore to the exploitation of offshore stocks, there to replace foreign effort that may be forced out by restrictive regulation. This provides an opportunity to solve the problem of Canada's east coast fishing industry once and for all. Unfortunately the pressures of unemployment in the east coast economy remain exceedingly high. There are strong indications that politicians will not be able to resist the temptation to use the opportunities for increased offshore fishing not so much to divert excess labor resources from the inshore fishery, as to accommodate larger number of fishermen at continuing low levels of income and/or high rates of subsidization.

IV. Perverse Effects

While the purpose of extended fisheries jurisdiction has been to give better protection to the fisheries of coastal states, it may have an opposite effect in several instances that affect Canada. Potentially most serious—though not now acute—is the threat that distant water vessels displaced from 200-mile zones around the world will divert their attention to high-seas salmon stocks of North American origin. There could also be increased foreign fishing pressure on other transboundary stocks, such as those straddling the 200-mile limit on the Grand Banks of Newfoundland. Some of Canada's traditional coastal fisheries of high value, such as those for Pacific halibut and Atlantic scallops, are conducted on grounds that may fall for a large part within the *U.S.* 200-mile zone. The magnitude of the potential loss in these fisheries to Canada depends on the outcome of negotiations with the United States on reciprocal fishing privileges and disputed boundary determinations.

There is also a trade diversion threat to the Canadian fishing industry from the 200-mile limit. Revival of a large New England groundfish industry is being promoted, based on reservation of *U.S.* 200-mile stocks for use of the domestic fleet. The principal obstacle remaining to such a development is importation of cheap Canadian groundfish. A concerted effort is underway to induce the *U.S.* government to restrict Canadian imports, with a principal argument being that Canadian fishing industry subsidies justify the imposition of countervailing duties. As groundfish exports to the United States are the mainstay of Canada's economically vulnerable Atlantic coast fishing industry, the consequences of *U.S.* import restrictions could be quite severe.

V. Conclusion

The 200-mile limit is proving generally beneficial in the regeneration of resource rents and the restoration of overfished coastal stocks. However, the territorial definition of fisheries jurisdiction has limited the effectiveness of the new resource management regimes, with uncertain effects on transboundary stocks. And the reduced access of distant water fishing fleets to coastal zones may lead to further overexploitation of stocks in the remaining high seas.

The distribution of the enhanced benefits flowing from the new 200-mile management regimes naturally favors the coastal states. Among them the allocation of benefits will be affected by interzone boundary determinations and by trade diversions resulting from international shifts in primary sector activities.

Canada's experience is illustrative both of the opportunities available for enhanced economic returns to coastal states, and of the residual and perverse problems associated with extended jurisdiction. The pressure for excessive effort in the inshore fishery has not abated, though the country is now able to reserve larger stocks to accommodate that pressure. With the longer-run prospect of a nearly exclusively domestic fishery in the Canadian 200-mile zone, a beneficial short-run accommodation has been arranged with foreign fishing fleets.

On balance, extended jurisdiction appears to be a step in the direction of more rational resource use. It may foster better habits in resource management and a better understanding of management needs. In turn, this could help to promote cooperation among nations in the joint management of the

remaining international common property resources.

REFERENCES

F. T. Christy, Jr., "Property Rights in the World Ocean," *Natur. Resources J.*, Oct. 1975, *15*, 695–712.

P. Copes, (1972a) "Factor Rents, Sole Ownership and the Optimum Level of Fisheries Exploitation," *Manchester Sch. Econ. Soc. Stud.*, June 1972, *40*, 145–63.

———, (1972b) *International Fishery-Resource Management: A Position for Canada*, Canada Department of the Environment, Ottawa, 1972, rev.

———, "The Law of the Sea and Management of Anadromous Fish Stocks," *Ocean Develop. Int. Law J.*, No. 3, 1977, *4*, 233–59.

———, "Canada's Atlantic Coast Fisheries: Policy Development and the Impact of Extended Jurisdiction," *Can. Publ. Pol.*, Spring 1978, *4*, 155–71.

H. S. Gordon, "The Economic Theory of a Common Property Resource: The Fishery," *J. Polit. Econ.*, Apr. 1954, *62*, 124–42.

A. Gotlieb and C. Dalfen, "National Jurisdiction and International Responsibility: New Canadian Approaches to International Law," *Amer. J. Int. Law*, Apr. 1973, *67*, 229–58.

Canada Department of the Environment, *Policy for Canada's Commercial Fisheries*, Ottawa 1976.

United Nations, *Informal Composite Negotiating Text*, Third Conference on the Law of the Sea, A/CONF.62/WP.10, New York 1977.

The Economics of the Eastern Bloc Ocean Policy

By VLADIMIR KACZYNSKI*

As a background to the analysis of the development of marine resources of the Soviet bloc countries, it seems appropriate to compare fishery activities of the Eastern nations with world tendencies in the utilization of marine-living resources. Table 1 indicates the following trends in the socialist countries use of ocean resources in recent years.

1) During the first half of the decade of the 1970's, the Soviet bloc rapidly increased its share in the total world harvest of marine organisms.

2) The rate of growth of the catch volume in socialist countries was many times higher than that registered during the same period by world fisheries.

3) Beginning in 1976, Soviet bloc fishing nations showed a sharp slowdown in the rate of increase of their catch. Some fisheries (for example, Poland) even registered a decrease in harvest compared to 1975. This was probably the turning point in fisheries development in the Eastern nations.

One of the most important reasons for the rapid growth of the Soviet bloc's long-range operations is that Eastern countries have based their expansion on resources of high abundance, even if they are not of highest value. Large concentrations of commercial species underutilized by coastal nations are presently the main targets of Eastern fleets (see Jan Elwertowski). Soviet bloc fisheries have shown a high level of technological flexibility and adaptability to the various conditions of resource use. This has enabled them to make frequent changes in ocean areas and species exploited, while at the same time maintaining high production rates in their harvesting process (see A. T. Pruter). The ability to respond to the variety of challenges of ocean resource use is in turn based on the well-developed maritime economies of some Eastern countries, particularly of the shipbuilding industries in Poland, East Germany, and the Soviet Union. Soviet fishing potential experienced tremendous growth based principally on supplies obtainable from the German Democratic Republic (*GDR*) and Polish shipyards. Bulgaria and Rumania also developed their ocean fisheries with ships built mainly by the *GDR* and Poland.

According to the *Lloyd's Register of Shipping Statistical Tables* six socialist fishing nations (Bulgaria, Rumania, Cuba, *GDR*, Poland, and the *USSR*) owned 60 percent of the total world fishing fleet tonnage (vessels over 100 GRT). The Eastern bloc's fleet is also growing more quickly than the tonnage of the world fleet. In the Soviet Union, the most characteristic feature of fishing fleet development is the increasing number of large factory trawlers over 2,000 GRT each. In 1977, this part of the Soviet fleet constituted about 64 percent of the total Russian fishing potential. In 1977, the total number of factory supertrawlers (over 2,000 GRT) of the six Soviet bloc fishing countries reached 932 vessels. This fleet grew during the period from 1975 to 1977 by 215 vessels and it is the group which demonstrates the fastest growth rate in Eastern fishing fleets.

Large factory trawlers are necessary for developing super-deep-water and pelagic fisheries. In open ocean areas and in Antarctic operations, the difficult hydro-meteorological conditions and the long distances from home ports require high seaworthiness, a high degree of autonomy of navigation, and high engine power. In their ocean fishery activities, Eastern countries are able to harvest any type of marine-living resources with the full utilization of their internal economic and social factors, and without the necessity of imports or larger hard currency outlays. Indeed, exports of a part of the fish production generates currencies which in turn can be devoted to finance emergency fish imports or to pay license fees for fishery activities within

*NORFISH, Institute of Marine Studies, University of Washington.

TABLE 1—WORLD AND SOVIET BLOC TOTAL CATCH IN 1969–76

Years	World (1)	World (2)	USSR (1)	USSR (2)	Poland (1)	Poland (2)	Other Eastern European Countries (1)	Other Eastern European Countries (2)
1969–71ª	67.9	–	7.0	–	0.465	–	0.519	–
1972	66.2	−2.4	7.8	+10.8	0.544	+17.0	0.576	+110
1973	66.8	+0.9	8.6	+10.6	0.580	+6.6	0.617	+71
1974	70.4	+5.4	9.2	+7.8	0.679	+17.1	0.657	+65
1975	69.6	−1.2	9.9	+7.7	0.801	+18.0	0.721	+97
1976	72.1	+3.6	10.0	+0.9	0.775	−3.3	0.731	+14
1977	–	–	9.6	−4.0	0 664	−14.0	–	–

Note: Column (1) = millions of metric tons; column (2) = increase in percent
Sources: F.A.O. Yearbook of Fishery Statistics, 1969–76 and National Marine Fisheries Service (*NOAA*).
ªYear average.

the 200-mile economic zones.

However, the magnitude of the Eastern oceangoing fleet reflects a massive investment and tremendous commitment to the exploitation of ocean fisheries. With the extension of national jurisdiction to the traditional areas of the Eastern fleet's operation, this commitment will be seriously affected. In socialist countries, the rapid build up of ocean harvesting potential began with the development of nearby coastal resources, but these quickly became insufficient as the market demand for fish grew. During the last thirty years, Eastern countries have developed full-scale commercial fisheries specialized in harvesting marine resources that are distributed over continental shelves adjacent to foreign coasts. On the eve of widespread implementation of the 200-mile economic zone, Soviet bloc countries took about 50–70 percent of their marine resources in fishing grounds located from 2,000–9,000 miles from their home ports. Southern ocean fishing grounds, which are presently being exploited, are located as far as 14,000 miles from Baltic harbors. One result of this distant fishing has been an increase in harvesting costs.

Although technological progress in Eastern fisheries has contributed greatly to the expansion of these nations in all ocean areas, nevertheless, the consumption of energy and outlays related to these changes have increased rapidly. Fuel consumption became the single most important cost item in the operation of the ocean fishing fleets. In just the last ten years (1965–75), energy costs have increased more than six times and they presently constitute over 20 percent of the total harvesting outlays. The growing scarcity of the ocean resource base of Eastern fisheries has resulted in the rapid growth of extraction costs (over 2.5 times during 1965–75), with energy costs being the most important and constantly growing component of the total harvesting outlays (see Zofia Szymanski).

In addition, the structure of catch has suffered undesirable changes from the highly valuable and demanded species (cod, halibut, red fish, herring, etc.) to the new species of small dimension, difficult processing, and low market value (mackerel, anchovies, capelin, dogfish, etc.). Finally, the strategy of constantly shifting the distant-water fleets from one area to another is gradually losing its positive economic effect. Legal limitations, depletion of the traditionally harvested species, and other restrictions have sharply reduced fishing opportunities of Eastern fleets.

One of the most significant features of the Soviet bloc's fishery industry is that the foundations for its modern ocean harvesting potential were laid on a highly restricted resource base. Only recently has the importance of this become clear. Although many other nations had limited coastal marine-life resources at their disposal, the peculiar economic and political context in which Soviet bloc ocean harvesting potential developed made this particular aspect one of

crucial importance for the future of its ocean expansion. This means that practically all Eastern states are heavily dependent on the marine-life resource base which is presently enclosed by the national jurisdictions of coastal states. The most important part of these resources is actually under the jurisdiction of developed Western nations—mainly the United States, Canada, European Economic Community countries, Australia, and New Zealand. Eastern fleets had been harvesting these resources for many years, but are now being forced to suspend their operations in some areas (the North Sea, Norwegian Sea, and adjacent waters) or to reduce the catch sharply in others (*U.S.*, Canadian, and West African waters).

Alternate sources of marine-animal supplies cannot immediately be developed by individual Eastern countries because this task would require many difficult and expensive surveys and investigations in the still unrestricted coastal waters and in the open ocean. Consequently, with the partial exception of Poland, Eastern countries will be heavily dependent upon Soviet research data concerning potential ocean resources and on its experience and know-how in utilization of unconventional resources (krill, open ocean and super-deep-water fisheries, etc.). Furthermore, they may rely upon the highly developed Soviet support fleet potential and on Soviet coastal zone resources. In the final analysis, the Soviet Union will be able to maintain its leading role in all Eastern ocean policy decisions and in the future development of the marine-life resources by the socialist nations.

According to Soviet scholars, the ocean-living resources should be exploited on an equal basis by all nations to the maximum sustainable extent without endangering the species. This theory has been developed because the Soviet Union, more than other maritime powers, relies so greatly on access to waters lying off the shores of other states. Consequently, claims of the coastal states to their adjacent marine-life resources are acceptable for the Soviet bloc provided that its countries will be allowed to continue their activities in the enclosed areas on the basis of full utilization of the coastal fishery resources. During the last few years, the Soviet Union and other Eastern states have witnessed the growth of optimum yield application in management of coastal resources which, in the eyes of some Eastern writers (see Jozef Popiel) is designed to further reduce the fishing opportunities of foreign users.

The economic efficiency of ocean resources use is presently being analyzed in Eastern states by applying an input-output method and by evaluating the contribution of ocean-living resources to the Eastern economies as compared to that of other protein producing sectors (see Andrzej Niegolewski). This method, commonly known in socialist literature as a balance of intersectorial flow of goods and services has, however, been criticized lately by Nikolaj Fiedorenko and others as inefficient in determining the optimal allocation of the resources among different sectors of the national economy. It is possible to determine the "optimal" intersectoral proportions with this method, but there is no guarantee that within these sectors the most effective alternatives of production development were chosen. For example, in the agriculture of Poland, primitive technologies still prevail which contribute to lowering the level of social labor efficiency and to increasing social costs of food production (see Augustyn Woś). Consequently, the question arises as to whether underdeveloped agriculture can serve as a suitable point of reference in evaluation of the ocean harvesting and seafood producing activities?

In Soviet bloc fisheries, major decisions related to expansion and investments in the ocean harvesting industry are made at the highest government levels. The state, as sole sponsor of ocean technology development and as owner of a large sea harvesting potential, is highly interested in the quickest and easiest solution of the food supply deficit in its home market (see Alexander Ishkov).

In planning the future development of ocean expansion during the era of unrestricted access to the ocean resources, Eastern planners always had assumed the most advantageous biological and technological alterna-

tives, that is, easy access to any resources and direct interrelationship between increase of harvesting capability and growth of the harvest volume. These principles did not substantially change even during the period when the catch limitations went into effect in all oceans. In order to increase the fish catch in the state yearly plans, it was enough to increase the fishing potential or assume better utilization of disposable fleets.

However, during the last few years, increased investments in distant-water fishing fleet development has not resulted in a proportional increase in catch. This alarming tendency is shown in Figure 1 which is based upon data in Table 1 and *Lloyd's Register of Shipping Data.* In order to achieve a slight total catch increase, it was necessary to pay a higher and higher price. For example, in Poland, the distant-water fishing fleet capacity increased about 25 percent during the 1975–77 period, while during the same period of time the volume of catch decreased 17 percent. Similar tendencies can be observed in the development of fleet capacity and additional catch volume in the Soviet Union and other Eastern fishing countries. The present state of Eastern fisheries, in which decreasing growth rates in catch are accompanied by disproportional increases in costs of ocean fisheries, has become a heavy economic burden for many Eastern governments. They are now forced to support these activities with growing subsidies. Without state support, long-range ocean activities would be impossible in these countries or would have to be totally restructured.

The present expansion of state-owned ocean fisheries in Eastern nations creates a situation in which the government meets production deficits of one group of commodities (bread, milk, fish, etc.) by charging

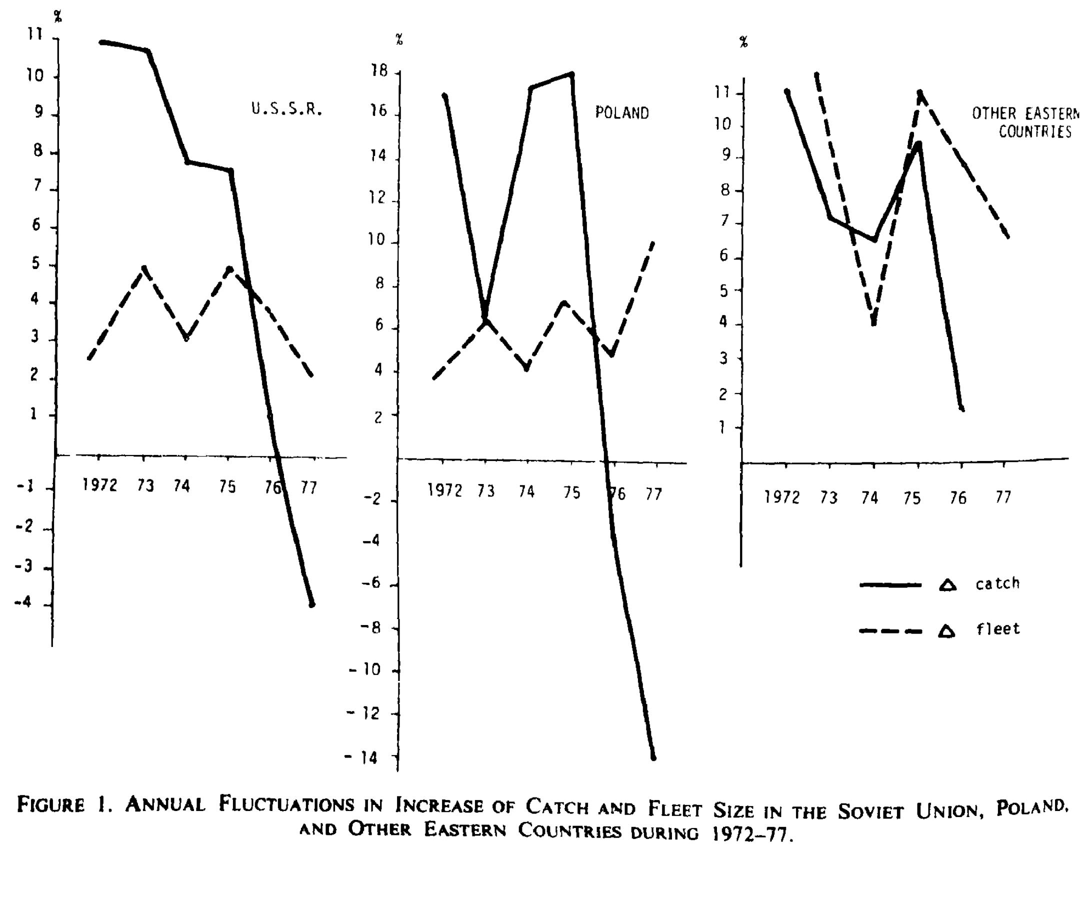


FIGURE 1. ANNUAL FLUCTUATIONS IN INCREASE OF CATCH AND FLEET SIZE IN THE SOVIET UNION, POLAND, AND OTHER EASTERN COUNTRIES DURING 1972–77.

incommensurable taxes for others (clothes, furniture, cars, etc.). As a result, the economic status of Eastern ocean harvesting activities constitutes an integral part of an acute market disequilibrium, with multiple consequences for the whole economy of the Eastern bloc countries.

REFERENCES

J. Elwertowski, "Sytuacja Swiatowego Rybolowstwa" ("Situation of the World Fisheries"), *Technika i Gospodarka Morska* (*Sea Technology and Economy*), Sept. 1977, *27*, 531–34.

N. P. Fiedorenko, "O razwitii sistema modeli v planirovanii narodnogo khozjajstva" ("On the Development of the Models System in National Economy Planning"), *Ekonomika i Matiematiczekije Mietody*, May 1977, *13*.

A. Ishkov, "Morski Przemysl Rybny Zwiazku Radzieckiego" ("Sea Fishery Industry of the Soviet Union"), *Technika i Gospodarka Morska* (*Sea Technology and Economy*), Nov. 1977, *26*, 641–43.

A. Niegolewski, "Przemysl rybny w bilansie przeptywow miedzygaleziowych" ("Fishing Industry in the Intersectorial Flows Balance"), *Bull. Sea Fisheries Instit.*, June 1976, *6/38*, 5–9.

J. Popiel, "Present Day Problems in Polish and World Fisheries," *Polish Maritime News*, Sept. 1977, *9*, 19–20.

A. T. Pruter, "Soviet Fisheries for Bottomfish and Herring off the Pacific and Bering Sea Coasts of the United States," *Marine Fisheries Rev.*, Dec. 1976, *38*, 1–15.

N. P. Sysoew, *Economics of the Soviet Fishing Industry*, Moskva 1970.

Z. Szymanska, "Skutki kryzysu paliwowego w polskim rybotowstwie morskim" ("The Impact of the Energy Crisis in Polish Ocean Fisheries"), *Technika i Gospodarka Morska* (*Sea Technology and Economy*), Dec. 1976, *23*, 763–64.

E. Wisniewski, "Aktualne i przyszlosciowe zadania gospodarki rybnej" ("The Present and Future Tasks of the Fishing Industry"), *Technika i Gospodarka Morska* (*Sea Technology and Economy*), Mar. 1977, *27*, 129–31.

Augustyn Woś, *Zwiazki rolnictwa z gospodarka narodowa* (*Inter-Relationship of the Agriculture with the National Economy*), Warszawa 1975.

United Nations, F.A.O. Yearbook of Fishery Statistics, Rome 1969–76.

Lloyd's Register of Shipping Statistical Tables, London 1970–77.

National Marine Fisheries Service, "Latest Developments in World Fisheries," Part II, IFR-78-100, Washington, June 1978.

Marine Resources: The Economics of U.S. Ocean Policy

By JAMES A. CRUTCHFIELD*

There appears to be a tendency to ascribe something romantic to minerals or anything else that comes out of the ocean—especially if you do not have to do the work yourself. But there is nothing unique about marine resources. To the extent that the sea is capable of producing minerals, it will produce them to man's benefit when, and only when, demand and cost factors for land-based production of the same minerals are such that we do better by getting it from the sea than elsewhere. The proper way to evaluate the present and prospective potential of the sea is to set it within the framework of the supply of resources and the demands on those resources from all sources. The present value of most mineral resources in the marine environment (with the obvious exception of oil and gas) is at or near zero. Their significance lies in the ceiling they impose on future increases in prices for the same minerals from conventional land sources.

I. Minor Resources of the Sea: Energy and Fresh Water

Man has learned to produce useful power from tidal sources from several places at a cost about four or five times greater than the next best alternative. This, combined with overwhelming problems of conflict with competing users, suggests that the United States will find little of interest in this source of energy. Perhaps more promising is the possibility of obtaining power from temperature gradients. Under some conditions useful amounts of power may well be produced by this method. While the prospect of really massive production of energy from the sea is not on the horizon at the moment, the technology to produce relatively small but usable amounts of electric power may well be. The technology very likely will come from the sophisticated science and engineering of the developed countries, but the underdeveloped countries will probably be the first to use that technology in a practical way. Production of power from some combination of wind and thermal gradient or even wave action generators in small quantities in isolated areas, particularly in the tropics, would be highly attractive.

Fresh water seems another major resource of the sea; in our time desalinated water will not be feasible at prices approaching the incremental cost of water from more conventional surface sources, at least as far as the United States generally is concerned. There are many parts of the world, however, in which desalination is already an economically viable source of water for direct human consumption, though not for irrigation. And in some parts of the United States newly developed technologies offer the promise of additional water supplies at a fairly high cost but in small increments: an investment option that might be very useful to small coastal communities where the only other alternative would be a fifty-year commitment to a major river basin transfer system which would be underutilized for the first thirty or forty years. It is also worth noting that pure water can be produced from brackish water at a much lower cost than from the sea. A system using brackish groundwater below the Imperial Valley of California, for example, could deliver processed water to an existing distribution system with present techniques at attractive prices, if coupled with a thermal power generating operation.

II. Mineral Resources of the Sea

The oceanic exploitation of hard rock minerals (compacted minerals) would require a true mining operation. At present there are none; we have neither the means to locate nor the technology to recover and process any

*University of Washington.

compacted, hard rock mineral at the present time No known authority feels that it is even remotely in prospect.

Regarding dissolved minerals, their concentration in sea water is so low that there is no interest in recovering any of them except in the long-established industries producing salt, magnesium, and bromine. Of the three, salt is by far the most important; about 30 percent of the world's supply is produced from the sea. Until recently, the United States had been producing nearly a million tons of magnesium per year from sea water. Currently, however, such operations are not competitive with high-grade magnesium deposits on land. The same is true for bromine; virtually all bromine is now produced from land-based sources.

Unconsolidated minerals represent greater prospects. The most important current operations, by a considerable margin, consist of inshore dredging of sand and gravel. The United Kingdom, United States, Japan, and a number of other countries continue to recover substantial amounts of sand and gravel from the sea. The best estimate of the dollar value of annual production is about $80 million (1976). Operators in several countries are also getting very small amounts of alluvial minerals by dredging. These include tin (the most important in value), iron, aluminum, and zircon.

Prospective future enterprises in unconsolidated minerals include phosphates from phosphorite nodules and ferro-manganese nodules. Regarding the former, land based supplies of phosphates for fertilizer will continue to be of more uniform and better quality and lower priced than any possible phosphate production from the ocean for at least the next three or four decades. Over the very long run, however, ocean sources provide a comfortable future reserve. In contrast, the exploitation of ferro-manganese nodules has generated great interest and debate. The nodules referred to are consolidated, bonded minerals ranging from rice size to three or four pound chunks. In general, however, they run from walnut to potato size. They are found in enormous profusion scattered over the ocean floor in most parts of the world, with heaviest concentrations in the central Pacific Ocean at depths ranging from about 12,000–20,000 feet. The nodules are composed of a large number of chemical elements, four of which (copper, nickel, cobalt, and manganese) are highly interesting to man, but the actual metallic content varies widely.

All four metals are important to any industrial economy; they also happen to be metals which the United States imports. In spite of the term "manganese nodules," it is the copper and nickel (and, to a lesser extent, cobalt) in which mining companies are really interested. The quantities of these minerals available in the sea are measured in trillions of tons. Since the annual increment to the nodules is probably on the order of 10 million tons per year, they are, in economic terms, more closely akin to exhaustible mineral reserves than to renewable resources.

Capital requirements for utilization of manganese nodules are large (in the range of $.5 billion if the essential investment in expensive processing plants is included). All of the companies that appear to be planning commercial production within the next five years are international consortia of firms representing three to twelve industrial nations: partly to spread the risk, partly to develop international support for deep sea mining ventures in an area where legal title is still very uncertain, and partly to assure adequate financing for the highly risky ventures.

The problem of an adequate legal framework, about which American companies have expressed deep concern, is still unresolved, despite long rancorous arguments in a series of Law of the Sea Conferences. In general, there are two conflicting positions. The United States, Germany, and Japan, among other industrial nations, argue that an interim arrangement under which secure title to seabed mining tracts could be obtained is essential, but they still pay lip service to the idea that the Law of the Sea Conference should ultimately produce a multilateral international framework. (The private companies involved argue that this framework may be years in coming and that we should be getting out the minerals now.)

The conflicting point of view, being pushed largely by developing nations, is that the deep sea bed is the heritage of mankind, not the

preserve of a few technically proficient nations. They see no reason why the developed nations now technically capable of operating in the open sea should stake out all the promising areas before anyone else is in a position to claim part of the benefits. Neither side appears to give more than passing attention to the role of international trade in minerals and its effect on the distribution of benefits and costs from viable nodule recovery operations.

American industry has been pushing very hard for an interim *domestic* policy which would say, in effect, that American firms are free to go into the open ocean despite the absence of any international agreement to secure tenure for the tracts that they are licensed to exploit by the *U.S.* government. If any management regime is later created by the Law of the Sea Conference, and if the United States becomes a party to that agreement, it is argued that the *U.S.* government should insure the companies against any losses they might suffer as a result of being restricted by or forced to pay taxes or fees to a new international entity.

However, there is real doubt about the wisdom of pushing ahead on a forced-draft basis from the standpoint of our own national interest. There is a good deal of basic logic in the position of the developing countries, and an even larger amount of emotional support for that position. If the United States and a few other developed countries unilaterally create a situation of *de facto* property rights, in the face of international disapproval, they will polarize international opinion in the United Nations and elsewhere to a degree that might *really* pose a threat to our access to minerals. The *U.S.* stake in getting some copper, cobalt, and nickel from the sea on a preferred basis is pitifully small compared to its interest in maintaining an orderly trading community of nations.

III. Gas and Oil

Petroleum and natural gas are critically important to the energy situation worldwide. Both are being produced offshore along the coasts of some fifty nations at the present time, and exploratory drilling is going on off the coast of more than forty others.

The United States seems to have taken a firm position that it is in the national interest to recover our own offshore oil at top speed. However, the United States still has substantial offshore reserves. Assuming that the real price of oil will be increasing steadily over the intermediate and longer term, is there any reason to believe that we would not be better off in the long run to use other people's oil as long as they will sell it to us? It *is* essential to define as accurately as possible the resources that will still be under our own physical and political control, in order to assess the benefits of having gas and oil when the external sources begin to become very costly—as they must in time. But there is no obvious reason to force early production.

There is clearly a conflict—real, not imagined—between the well-being of a private oil industry and the public interest of the American people. From the standpoint of the oil industry, the method chosen to lease oil-bearing land offshore (bonus bidding plus royalties) guarantees that once oil is discovered in an outer continental shelf area, and bidding has been opened for the right to explore and exploit, a financial commitment of really major proportions has already been made.

For example, some companies have laid out, in the Gulf of Mexico and off Santa Barbara, as much as a \$.25 billion for a single lease; and this is only a "hunting license." At a minimal opportunity cost of 10 percent the successful bidders are paying out hundreds of thousands of dollars a day for lease rights only. Thus, the pressure to produce is tremendous; but it is not clear that it is equally in the interest of the United States (or the larger global interest of the world community) that we rush out to produce offshore petroleum as quickly as we can.

Prudhoe Bay provides an excellent example of "haste makes social waste." After the headlong rush to develop North Slope oil resources, growth in demand for petroleum products has been slowed by price increases and other measures. The North Slope oil coming out of the end of the pipe in very large

quantities has generated substantial surpluses, and there is no place to put them except through the Panama Canal. But where else can the tankers go? The law says that North Slope oil must go to an American refinery. But there is nowhere on the West Coast to sell it. There are no pipelines to carry it East. Further, it takes at least four years to get a terminal and pipeline on line, by the time all the environmental impact statements and physical construction are completed. We will probably end up selling the oil to Japan—which probably would have been the most economic thing to do in the first place, though it makes a mockery of all the arguments for frantic haste in development.

Environmental aspects of petroleum are crucial. Every stage—producing, transporting, transferring to shore-based establishments, processing, and using oil—creates environmental hazards. There is no way that man can produce and transport oil from a marine environment that is not many times more hazardous to living marine life and to other elements of man's well-being than if it were produced on land. Yet we have made only the most slipshod kinds of preparations for dealing with such dangers. The Bureau of Land Management is now busily doing impact studies of offshore drilling off the Pacific Coast states and Alaska, while engaging simultaneously in leasing these territories.

IV. Living Resources of the Sea

In 1946 world production of fish was about 20 million metric tons. By the late 1960's, that figure had risen to about 70 million metric tons: an annual rate of increase of around 6 percent a year, and a spectacular increase in the supply of protein food available to the world. From 1970–76 there has been no further increase in the world landings. Yet in this period the increase in fishing capacity, in tonnage and efficiency, has been tremendous. The whole East European community, the Taiwanese, the Koreans, and others have been putting deep water, highly productive fishing equipment in every major ocean fishing area.

Even more alarming is the clear evidence the output of food fish has been declining since about 1950, and the growth of total fish production and the relative stability in the last few years has been almost entirely the result of an enormous increase in industrial fish production for meal and oil. Most has come from a handful of major fish meal operations; Peru is the principal one, and South Africa and Norway have also been important factors. In fact, then, the production of fish for direct human consumption has not been increasing over the last twenty years. The industry has been moving to lower valued species in order to maintain aggregate total production. Coupled with slowly rising demand, this has led to sharp increases in the *real* price of fish.

V. Future Prospects for the Fishing Industry

In the developed nations, the income elasticity of demand for fish is relatively low; per capita consumption has hardly changed over the last forty or fifty years. Hence, in these nations demand for fish per se is increasing only at a rate roughly approximating population increases. Japan is a notable exception. However, the demand for services associated with fish is highly income elastic (as is true of most foodstuffs). In the developing areas, many of which are dependent on fish as a major source of protein food, demand is highly income elastic and in addition population growth rates are very high. On a world basis, then, the aggregate demand for fish is expected to grow at a rate substantially greater than man's short-run capacity to expand production.

It is still possible to generate a good deal more useful protein out of the sea than our present markets, technology, and tastes would indicate. But the limit, beyond which further expansion becomes unattractive economically, is creeping up much faster than we had thought. One good example illustrates the point. At one time the fishery community was terribly excited about the prospects of food fish production in the Arabian Gulf. The initial Indian Ocean study conducted some time ago suggested very heavy concentrations

of fish in the area on the evidence of high basic biological productivity. When the Norwegians ran test cruises through the area they found that there was indeed a tremendous quantity of fish; but most of it consisted of mesopelagic species that nobody knows how to catch efficiently or process into marketable form. Instead of getting 2 or 3 million *marketable* metric tons from the area, the likelihood is something like 3/4 of a million metric tons—if we are lucky.

Thus, the world has been facing and will continue to generate severe conflicts in fisheries. The prospect of constantly rising demand for protein food and much tougher physical and therefore economic limitations on productivity than had been assumed in the past adds up to an increasingly serious basis for conflict as time goes on.

How well has man utilized living resources of the sea?

1) Have we actually exploited fishery resources at the proper rate to maintain productivity of the stocks?

2) Have we harvested them efficiently, so that any given rate of harvest can be conducted as cheaply as possible?

3) Have we provided inducements for fishermen to develop still better ways of harvesting?

4) Have we provided for any kind of orderly allocation of the fish catch of the world, in terms of both food and the incomes generated from the activity, among the countries and the people who participate?

Most resource economists would agree that the record of man's utilization of the living resources of the sea ranges from very poor to awful. We have failed signally to deal with the fact that the common property status of ocean resources—the inability to establish exclusive property rights in living marine resources—generates constant pressure toward overcapacity. In some cases depletion or even destruction of whole populations has occurred, and in virtually every fishery the industry uses far more gear and labor than is actually required to harvest any given level of catch. The number of examples is legion. Ten or twelve years ago a working group estimated that the catch taken from the Northeast and Northwest Atlantic—one of the great fishing areas of the world—could be taken with about 33 percent less effort than was actually being exerted, and that after a short period of lower catches, as a natural result of the reduced effort, the subsequent increase in the average size of the fish would have produced catches 3–5 percent higher than those being taken. The degree of overcapacity was estimated at that time to impose a dead weight burden of needless cost of $100 million a year. At today's prices and capacity it would be more like $2 or 3 billion a year. The same general tendency has developed on a worldwide basis throughout the fishing community.

Had the resources available been managed on a more rational basis, even within the limits of scientific knowledge that we are bound to struggle with, the total catch figure today might well be 90–100 million metric tons rather than the 70 million mentioned above. Living marine resources are not being utilized fully or wisely, and nothing in the international arrangements that had been tried in the forty or fifty years prior to 200-mile legislation had really made any significant difference in the tendency brought on by common property to misallocate resources in the exploitation of fisheries. Hence, the pressure for the 200-mile limit: a second best solution, clearly, but obviously better than the management regime—or lack of it—that it has supplanted.

It is much too early to assess the potential for economic improvement in *U.S.* fisheries and their management as a result of the extension of *U.S.* control to 200 miles. On the positive side the Fishery Conservation and Management Act of 1976 recognizes explicitly the multiple objectives of fishery management, including economic efficiency, and it has created a vehicle, in the eight regional councils that may correct the divisive effect of state jurisdiction. Less promising is the evident pressure to load the councils with industry spokesmen rather than persons with a broader view of the public interest and the severe limitations placed on effective measures to control excessive inputs. The New England region has already seen a "Gold

Rush" tendency to supplant redundant foreign fishing capacity with redundant American vessels—hardly a step forward.

Most important, neither the individual councils nor executive or legislative leaders have reached a clear concensus on objectives. It is not at all clear that *U.S.* interests are best served by having all fish taken by *U.S.* flag vessels within the 200-mile zone. A more rational policy toward fees for foreign users of those resources might suggest that we are better off to "rent" substantial segments rather than subsidize early and costly American expansion.

In short, the Act preserves for the United States a number of vital options that would have been irretrievably lost under uncontrolled international fishing off our coasts. It remains to be seen whether those options are utilized to rationalize an otherwise uneconomic pattern of resource use.

The Outlook for Social Security

By A. HAEWORTH ROBERTSON*

The Social Security Amendments of 1977 (P.L. 95-216), passed by Congress on December 15, 1977 and signed by the President on December 20, 1977, made important revisions in the Social Security Program. These amendments are generally considered to be the most significant Social Security legislation since 1972, and possibly since 1950.

The Social Security Administration has prepared projections of income and outgo for the Social Security trust funds taking these amendments into account. The assumptions and methodology used in making the projections are the same as those set forth in recent annual reports of the Board of Trustees on the Old-Age and Survivors Insurance, Disability Insurance, and Hospital Insurance (*OASDHI*) programs, and reference should be made to these reports for a more thorough description of the subject of financing. To facilitate the presentation of long-range cost estimates, expenditures are normally expressed as a percentage of the total income which is subject to the Social Security tax, that is, as a percentage of taxable payroll. This procedure avoids the difficulties created by the changing value of the dollar over time and allows a direct comparison to be made between projected expenditures and the tax rates scheduled in the law.

The total expenditures for the *OASDHI* programs are projected to rise from some 13 percent of taxable payroll in 1977 to approximately 16 percent in the year 2000 and 24 percent in the year 2025, remaining at about that level thereafter. The total tax rate, on the other hand, for employees and employers combined is scheduled to increase from 11.7 percent of taxable payroll in 1977 to only 15.3 percent by the year 1990, remaining level thereafter. Projected expenditures are higher than scheduled tax income beginning in the mid-1990's, with the deficit growing larger with time and becoming particularly large after the turn of the century. Therefore, the Social Security Amendments of 1977 did not solve all of the financial problems of the program. In particular, resolution of the following two potential problem areas was postponed: 1) With respect to the Old-Age, Survivors, and Disability Insurance program, the long-range financing problem beginning in the year 2011 caused by children of the post-World War II baby boom reaching age 65. 2) With respect to the Hospital Insurance program, the short-range financing problem caused by the continuing rapid escalation of hospital costs, and the long-range financing problem caused by the aging of the post-World War II generation.

Much was accomplished by the Social Security Amendments of 1977. Much more remains to be done.

I. What Changes Lie Ahead?

What kinds of change can be expected in the present law and in the behavior of the population covered by that law, not only because of these projected rising costs, but because of other factors? The following seven points seem to me to be reasonable expectations for the future (although I would not necessarily advocate all of these developments).

First, taxpayers must become accustomed to paying higher taxes for Social Security benefits unless benefits are reduced below current levels. It is just not possible to pay for the current Social Security Program with the taxes now being collected. If present financing methods are continued and the law remains unchanged, the employee and employer tax rates (for the *OASDHI* programs)

*Vice president, William M. Mercer, Incorporated.

must rise from their present level of 6.05 percent of taxable payroll to about 8 percent by the year 2000 and some 12 percent by the year 2025.

Second, it seems unlikely that the traditional financing methods will continue to be the sole source of tax revenue for the program. Taxpayers are increasingly asking what benefits they receive for their Social Security tax payments. As it becomes more evident that the relationship between taxes and benefits is tenuous for any given individual (that is, the program gives more emphasis to social adequacy than to individual equity), there will be increased resistance to payroll tax rate increases. This will probably result in the use of some form of nonpayroll tax (such as general revenues, a "value-added" tax, etc.) for at least one-third of Social Security expenditures sometime before the turn of the century. It seems unlikely that present Social Security payroll tax rates will be reduced significantly; the new form of taxation would represent additional taxes.

Third, all state and local government employees will eventually become participants in the Social Security Program. Perhaps some way will be found to make participation compulsory for state and local employees. Alternatively, if and when nonpayroll taxes are used to a significant degree to finance Social Security, state and local employees may insist on being covered by Social Security in order to receive their money's worth from their general taxes. Also, as the real costs of existing public employee retirement systems become more evident, there may be more of an inclination to reduce benefits under such systems and integrate them around the Social Security Program. These same factors will result in federal civil servants participating in Social Security and appropriately adjusting benefits payable under the civil service retirement system. Full participation by all state and local employees and federal civil servants would reduce the average long-range cost of the Social Security Program by about 2 percent (i.e., less than 1/2 percent of taxable payroll).

Fourth, beginning about twenty-five to thirty-five years from now employees will probably be working longer and retiring later. This will be a natural development as health and life expectancy improve, and as the growth in the work force slows because of the low fertility rates. For this to be feasible, present socioeconomic arrangements must be revised to make it easier for persons to continue working until advanced ages, perhaps in less strenuous jobs, part-time employment, etc. This development could lessen the financial problems of the Social Security Program during the next century since a later effective retirement age, other things being equal, is tantamount to a reduction in benefits. These cost savings would be negated, however, by further liberalizations of the retirement test or further increases in the delayed retirement credit.

Fifth, social and economic changes in the nation will result in substantial revision of the program. The changing role of the family unit, and of women; changing patterns in the incidence of work, education, and leisure throughout a person's lifetime; lengthening life expectancy and improved health in old age; an increased (or reduced) need to work in order to maintain the desired standard of living; all of these changes and more will require that drastic revisions be made in the benefit structure if the evolving economic security needs are to be satisfied appropriately. The net effect of all these changes will not necessarily be an increase in costs.

Sixth, if the nation experiences sustained inflation at relatively high levels, it is likely that the portion of an individual's economic security needs which can be met by the private sector will decrease over time; the needs must somehow be met; and the federal government (probably through the *OASDI* program) will be left as the only entity with the audacity to make unqualified promises to pay benefits 75 to 100 years in the future based upon indeterminable cost-of-living increases. Obviously, the cost of an expanded Social Security Program would be correspondingly increased.

Seventh, the Medicare program as well as the nation's entire health care system will be changed beyond recognition during the next twenty-five years. This will be the inevitable

result of diverse attempts to make more adequate health care available to the populace, but at the same time prevent total health care costs from continuing to rise as a percentage of the Gross National Product. In a reorganization of the Department of Health, Education, and Welfare announced on March 14, 1977, the administration of the Medicare program (formerly under the Social Security Administration) and the Medicaid program (formerly part of the Social and Rehabilitation Service) was placed under the authority of a newly formed Health Care Financing Administration, bringing the management of the nation's two largest health programs under one agency. These two programs will be reshaped in various ways and will probably evolve into a comprehensive national health insurance program.

II. Social Security's Two Largest Problems

It is unlikely that rational changes can be made in the Social Security Program so long as the present low level of understanding of the program persists. In the future, public understanding or misunderstanding will play a much more critical role in determining the shape of the program than it has in the past when the payroll tax was relatively low and when the taxpayer was in a less questioning frame of mind.

Therefore, I consider the most important problem confronting Social Security to be widespread lack of understanding of the program—the type and level of benefits it provides, the method of financing, the (tenuous) relationship between taxes paid and benefits received by an individual, etc. There is no excuse for permitting this lack of understanding to continue. Widespread understanding of the Social Security Program may result in a certain amount of trauma and even disruption, but even more disruption will result if the current misunderstanding is allowed to persist. If people understand Social Security, there is a much greater chance that the program will be modified to coincide with their desires and thus gain the public acceptance which is obviously necessary for a program which will pay benefits, and require tax collections, of over $2 trillion during the next 10 years.

I consider the second most important problem which Social Security must face to be the possibility of sustained high levels of inflation. Such inflation will result in a steady diminution of the role of private pensions and private savings with a consequent need for corresponding increases in Social Security benefits, further adding to long-range costs. It may result in an eventual conflict between the working and nonworking populations since it may not be possible to protect the nonworking population against the ravages of inflation except at the expense of the working population. Continued high levels of inflation will have other disruptive effects on the economy—probably more serious than many realize.

Of course, there are other problems with the Social Security Program. But if we can solve these two large problems, lack of understanding and inflation, all the rest of the problems will probably be manageable. If we cannot solve these two large problems, I submit that we are in for a long period of turmoil.

Disability Insurance

By PAUL N. VAN DE WATER*

The cost of the Social Security Disability Insurance (*DI*) Program has been growing very rapidly. Disability incidence rates (the number of new disability benefits awarded per insured worker of a given age and sex) doubled between 1965 and 1975, although they have remained stable since then. The number of disabled primary beneficiaries has risen from 1.2 million a decade ago to 2.9 million today. Real benefit costs have doubled in the past six years and will be about $13 billion in 1978.

While the number of disabled persons entering the *DI* rolls has been increasing rapidly, the number of persons leaving the rolls has been declining. Between 1967 and 1977 the rate of medical recovery or return to work fell from 3.3 to about 2 percent of the beneficiary population.

The Social Security Amendments of 1977 raised the *DI* payroll tax rate from 0.575 percent of taxable earnings for both employers and employees in 1977 to 0.775 percent in 1978, and 0.825 percent in 1981, an increase of almost one-half. Were it not for these tax increases, the disability insurance trust fund, which had been running deficits since 1975, would have been exhausted in late 1978. The 1977 tax increases and benefit adjustments, however, are now projected to keep the *DI* trust fund solvent through about the year 2020.

In this paper I present five propositions which attempt to explain this experience and explore its implications for public policy.

PROPOSITION 1: *The disability insurance program is a social insurance (non-means-tested) program designed to replace earnings lost when a person becomes unable to work.*

The disability insurance program first paid benefits in 1957, but the idea of a public program to replace earnings lost as a result of disability dates back to the beginning of the Social Security System. The hesitation to include disability insurance as a part of Social Security was largely due to the disastrous experience of private insurance carriers during the Great Depression. Prior to 1929, many insurance companies offered disability income benefits as a rider to life insurance policies. However, the massive unemployment of the 1930's resulted in many more claims than the carriers had expected, and the firms suffered great losses.

The fear of spiraling costs made the first steps in disability insurance very tentative ones. The definition of disability was—and still is—very restrictive. Not only must a person be unable by reason of an impairment to do his previous or customary work, but his impairments must be of such severity that he is unable—considering his age, education, and work experience—to perform any substantial paid job that exists in significant numbers in the national economy. Whether there are any vacancies for such jobs or whether the applicant would be hired if he applied is irrelevant. Earnings of more than a specified amount, now $280 per month, are considered evidence that a person is "capable of engaging in substantial gainful activity" and is therefore not disabled.

As an additional cost-saving feature, the technical or insured status requirements are more stringent for disability insurance than for old-age and survivors insurance (*OASI*). Not only must a worker have earned a certain number of quarters of coverage over the course of his career (as in *OASI*), but he must also demonstrate a recent attachment to the labor force, generally by having earned 20 quarters of coverage in the ten years immedi-

*U.S. Department of Health, Education, and Welfare. The opinions expressed in this paper are not necessarily those of the Department of Health, Education, and Welfare. I wish to express my indebtedness to Michael Barth, Patricia Dilley, Joseph F. Faber, Aaron Krute, Mark Miller, H. Elizabeth Peters, and Lawrence Thompson.

ately before becoming disabled. Finally, *DI* benefits are not payable until a person has been disabled for a full five months. (The 1974 Ways and Means Committee staff report provides a more complete description of the *DI* program. See also Mordechai E. Lando and Aaron Krute.)

PROPOSITION 2: *For reasons of administrative feasibility, the awarding of disability insurance benefits is and must be based on the presence of a serious medical impairment rather than on a finding that the particular individual is unable to work.*

To understand this point, one must comprehend the distinction between impairment and disability. Impairment is a clinical concept, which describes some physical or mental abnormality, such as diminished lung capacity, loss of limbs, or poor hearing. Disability is a behavioral concept—the inability to earn a living which may result from an impairment. Whether an impairment is disabling for a particular individual depends on nonmedical factors, such as his age, education, work experience, motivation, and state of the labor market (see Saad Z. Nagi).

In principle, the *DI* program pays benefits on the basis of disability—that is, inability to work. In practice, however, benefit awards depend on the actual absence of earnings and the presence of a specified impairment. The causal link between the impairment and the lack of earnings necessary for a finding of disability is *presumed*. A procedure of this kind is certainly necessary to produce reasonably replicable decisions in a system that handles almost 2 million claims per year. Moreover, the criteria that are applied are very stringent. One study has found that, even of those who apply for benefits and are turned down, four-fifths never return to sustained competitive employment.

The presumption of disability based on impairment, however, is still far from exact. In his seminal study, Nagi compared Social Security findings of disability with comprehensive evaluations by special clinical teams. The teams included doctors, social workers, occupational therapists, and vocational counselors. In three-quarters of the cases where the clinical team found the subject not fit for work, Social Security allowed the person's claim. Similarly, in three-quarters of the cases which the team found fit for work under normal conditions, the claim was denied. But where the team judged the person fit for work under special conditions, the split between allowance and denial was fifty-fifty.

PROPOSITION 3: *The administrative procedure for determining who is disabled causes the disability insurance program to be sensitive to changing economic and social conditions.*

As noted earlier, a given impairment may cause a person to become disabled, or to regard himself as disabled, but it need not do so. Calculations based on the Social Security Administration's 1972 *Survey of Health and Work Characteristics* show, for instance, that of people who have the requisite quarters of coverage to be insured for *DI* and who have the most severe impairments, almost half are working and have never applied for *DI* benefits. Members of this large reservoir of employed impaired people may be drawn onto the disability benefit rolls if the work ethic slackens, if the job market worsens, or if benefits become more attractive.

During the late 1960's and early 1970's, at least the latter two factors were operative. First, higher unemployment rates made it more difficult for impaired persons to find or hold jobs. Second, disability insurance replacement rates (the ratio of benefits to prior earnings) rose by over 20 percent from 1969 to 1973. In addition, disabled workers were made eligible for medicare benefits in 1972, effectively increasing replacement rates by another 25 percent.

As with most economic phenomena, there are varying estimates of the responsiveness of disability incidence rates to benefit levels. An unpublished time-series analysis by Mordechai E. Lando and Timothy Hopkins suggests that the elasticity of applications with respect to benefits is around 0.5 or 0.6. An individual cross-section study, also unpublished, by William G. Johnson and colleagues produces

similar estimates for women but zero elasticity for men. This is an area where more research is needed.

PROPOSITION 4: *The structure of benefits is the major policy tool available for affecting the size of the disability insurance program.*

Disability insurance benefits, like Social Security retirement benefits, are a function of the individual's average indexed earnings in employment covered by Social Security. The only difference is that in *DI* the average wage may be computed using a smaller number of years. The benefit formula, which is the same as in the retirement program, is weighted in favor of low-wage workers, who receive proportionately (but not absolutely) higher benefits than high-wage workers.

The primary benefit is computed as if the disabled worker were a 65-year old retiree. In addition, eligible dependents of the disabled person receive a benefit equal to 50 percent of the worker's benefit, subject to a specified maximum per family.

The resulting benefit payments are, on average, not particularly high. The average *DI* family benefit will be about $4,400 in 1978. We estimate that the median income of families with disabled workers in 1978 will be about $8,800, as compared to the median for all households of over $16,000. Nineteen percent of such families are living in poverty, as compared to only 9 percent of all families with a nonaged head. About 11 percent of disabled workers receive supplemental security income (*SSI*), a program for disabled persons who are not insured for *DI* or whose *DI* benefits are inadequate. (The basic *SSI* benefit is now $2,273 per year.)

Under the indexed benefit computation scheme that will go into effect in 1979 (see John Snee and Mary Ross), the average newly awarded *DI* family benefit will replace 48 percent of the worker's recent gross earnings and about 55 percent of recent earnings net of taxes and work-related expenses. Half of the disabled workers will have replacement rates of less than 50 percent of recent disposable earnings. But 16 percent will have wage replacement in excess of 80 percent, and 6 percent will have wage replacement over 100 percent. The instances of high replacement rates are due to the payment of dependents' benefits and, to a lesser extent, the weighted benefit formula.

If benefits are to be reduced to reduce costs and provide greater work incentives, there are three major approaches: 1) Limit total family benefits to a reasonable fraction, say, 70 or 80 percent, of the worker's previous earnings; 2) Reduce benefits across the board; or 3) Reduce the weighting in the benefit formula and/or the role of dependents' benefits.

While it is impossible to explore all the pros and cons of these options here, let me list four salient points. First, reducing benefit levels to discourage program participation by those who can work reduces the adequacy of benefits for those who are unable. This is the same tradeoff between adequacy and work incentives which is familiar in the context of welfare.

The second point is a corollary of the first. Any type of benefit reduction is bound to have an adverse impact on many people with very low incomes. Even though disability insurance is not targetted on the poor, it is still the single most important program for reducing poverty among the disabled. Were *DI* to disappear, about 60 percent of families with disabled workers would be living in poverty.

Third, even though disability insurance is a potent antipoverty program, it is relatively target inefficient. For the program as a whole, the estimated correlation between replacement rates and need (family income divided by the applicable poverty line) is -0.07, and even for men alone, a more homogeneous group, it is only -0.11. Thus, the presumption that people with low average indexed earnings or with eligible dependents are necessarily needy is quite costly. The amount saved by eliminating dependents' benefits, for instance, would be sufficient to raise the *SSI* guarantee for the disabled and virtually eliminate poverty among *DI* and disabled *SSI* recipients.

Fourth, the requirements for *DI* eligibility are sufficiently strict that two-thirds of the self-described severely disabled do not receive

benefits. Nevertheless, members of this group suffer substantial earnings loss due to health problems. Among severely disabled non-*DI* beneficiaries, the incidence of poverty is twice as great as among *DI* recipients.

PROPOSITION 5: *A strong case can be made for phasing out the disability insurance program and replacing it with an improved means-tested program for impaired people.*

Such an arrangement would have several advantages. Administration and public understanding would be improved through basing eligibility explicitly on impairment. Work incentives would be provided through the negative income tax mechanism, with earnings taxed at 50 percent—as in supplemental security income—or some other rate less than 100 percent; the complete loss of benefits which now occurs at earnings of $280 per month would be eliminated. The program would be financed by general revenues. And the program would concentrate scarce budget dollars on those most in need, rather than paying relatively large benefits to some seriously impaired people and nothing to others.

The disadvantages of this arrangement are twofold. First, income maintenance and wage replacement may be a package deal. The redistributional features of *DI* and *OASI* may be countenanced only because the middle class gets something too. On the other hand, the large program and the high payroll taxes that have resulted from this philosophy may have become counterproductive in maintaining public support.

Second, the private sector may not be capable of fully picking up the wage replacement function of disability insurance. The chief concern here is the inability, or at least unwillingness, of private insurers to provide indexed disability or retirement benefits. Similarly, private insurers may be unable to insure an event whose occurrence is so sensitive to macro-economic conditions. Private firms, however, may be better able than the government to make individualized determinations of disability.

Advocates of social insurance for the aged should not assume that a similar program is necessarily appropriate for dealing with disability. In the *OASI* and *SSI*-aged programs, eligibility is unambiguously based on age, and almost all aged persons are receiving either *OASI* or *SSI* benefits. Neither situation obtains in disability. The concept of disability is subjective; only one-third of the severely disabled are receiving *DI*, and only 60 percent are receiving any public income maintenance benefits (including *DI*). Using the social insurance approach to fill this large gap in disability protection would require making even more disability determinations and would greatly increase the costs of an already expensive program.

REFERENCES

W. G. Johnson et al., "A Study of the Determinants of Disability Insurance Applications," Health Studies Program, Syracuse Univ., May 31, 1978.

M. E. Lando and T. R. Hopkins, "Modeling Applications for Disability Insurance," paper presented at the ASSA Meetings, New York, Dec. 29, 1977.

——— **and A. Krute**, "Disability Insurance Program Issues and Research," *Soc. Sec Bull.*, Oct. 1976, *39*, 3–17.

Saad Z. Nagi, *Disability and Rehabilitation*, Columbus 1969.

J. Snee and M. Ross, "Social Security Amendments of 1977: Legislative History and Summary of Provisions," *Soc. Sec. Bull.*, Mar. 1978, *41*, 3–20.

Social Security Administration, *Survey of Health and Work Characteristics*, Washington 1972.

U.S. Congress, Committee on Ways and Means, *Committee Staff Report on the Disability Insurance Program*, Washington 1974.

Medicare: Its Financing and Future

By UWE E. REINHARDT*

The federal government's health insurance program for the aged—commonly referred to as Medicare—is now slightly over a decade old The program was enacted in 1965 as Title 18 of the Social Security Act. It went into effect on July 1, 1966.

I. The Structure of Medicare

Although it is customary to think of Medicare as one program, it actually consists of two completely distinct insurance plans: Part A of Title 18, the so-called Hospital Insurance (*HI*) plan, and Part B, the Supplementary Medical Insurance plan (*SMI*).

Part A of Title 18 covers hospital care, posthospital care in skilled nursing facilities, and home health care, including drugs and ancillary services rendered on an inpatient basis. It is financed through the Social Security payroll tax paid by employees and employers in equal parts. These contributions are mandatory. Part B of Title 18 is a voluntary supplementary medical program (*SMI*) covering physician services rendered on an inpatient or outpatient basis, hospital outpatient services, and certain home health care services not related to an inpatient episode. To qualify for Part B coverage, individuals must pay a monthly premium set at $3 in 1966 and at $8.20 currently. Eligible for Medicare coverage are all persons aged 65 and over receiving Social Security or railroad retirement benefits and, since 1973, any disabled person receiving cash benefits under Social Security and persons with end-stage renal disease.

II. Outlays under Medicare

Tables 1–4 present details on the secular growth of Medicare expenditures. The tables extend to 1976, the last year for which published data on all of the variables in the tables are available. Preliminary estimates indicate that Medicare expenditures during fiscal 1977 amounted to $20.8 billion, or 14.5 percent of the total national outlay of $142.6 billion on personal health care.

As is apparent from Table 1, Medicare expenditures have increased rapidly during its first decade, notably during the past few years. General price inflation has contributed much to this increase, as may be seen by comparing outlays in current and in constant dollars. Even after adjustment for growth in the number of enrollees and in the general price level, however, expenditures under the program show a marked upward trend. The growth in constant dollar per capita expenditures (lines A.3c and B.3c) can be explained in part by the relatively rapid increase in medical care prices, which in turn were driven upward by the Medicare program itself and by the growth of health insurance in general. Inclusion of the disabled and of persons with end-stage renal disease among the Medicare population also must have contributed to the increase.

Table 2 highlights the role of public sources in the financing of personal health care for the aged. According to these data, government sources constitute roughly 69 percent of total per capita expenditures in fiscal 1976, private health insurance 6 percent, and out-of-pocket payments by the aged (for drugs and other items not covered, as well as payments of deductibles and coinsurance) amounted to 26 percent.

Persons not familiar with Medicare may believe that the program covers most of the aged's expenditures on health care. Actually, the program itself accounts for less than half of these outlays. It paid for only 43 percent in fiscal 1976. Even that percentage overstates the net contributions of the program because Medicare expenditures are financed in part through *SMI* premiums on the aged. After adjustment for premium payments, Medi-

*Department of economics and Woodrow Wilson School, Princeton University. I would like to thank Ira Burney of the Health Care Financing Administration, Department of Health, Education, and Welfare for presenting this paper.

TABLE 1—ENROLLMENT IN AND TOTAL BENEFIT EXPENDITURES FOR THE MEDICARE PROGRAM, 1967–76

	1967	1973[a]	1976	Average Annual Growth Rate[c] 1967–73	1973–76
Total Benefit Expenditures under *OASDHI*[b]					
a. In billions of current dollars	$4.6	$9.6	$18.4	13.0	24.2
b. In billions of 1967 dollars	4.6	7.2	10.8	7.8	14.5
A. Hospital Insurance (Part A, Title 18)					
1. Number of Enrollees as of July 1 (millions)	19.5	23.3	25.3	3.0	2.8
2. Total Benefit Expenditures (billions)	3.4	7.1	13.3	13.1	23.3
3. Benefit Expenditure per Enrollee.					
a. In current dollars	172	302	527	9.8	20.4
b. In 1967 dollars	172	226	309	4.7	11.0
c. Deflated by price index for semiprivate room rate[d]	172	166	196	−0.6	5.7
B. Supplementary Medical Insurance (Part B)					
1. Number of Enrollees as of July 1 (millions)	17.9	22.5	24.6	3.9	3.0
2. Total Benefit Expenditures (billions)	1.2	2.5	5.1	13.3	26.7
3. Benefit Expenditures per Enrollee:					
a. In current dollars	67	112	206	8.9	22.5
b. In 1967 dollars	67	84	121	3.8	12.9
c. Deflated by price index for physician fees[d]	67	81	109	3.2	10.5

Sources: Data on expenditures and price indices: *Social Security Bulletin*, June 1978, Tables M-2, M-45 and M-46. data on enrollment: *Social Security Bulletin Statistical Supplement 1975*, Tables 138 and 139; *Health Insurance for the Aged and Disabled· Amounts Reimbursed by State and County 1976*, Table 1.1.1.

[a]As of July 1, 1973, coverage for disabled under age 65 began.

[b]Benefit expenditures from the Federal Hospital Insurance and Supplementary medical Insurance Trust Funds as reported by U.S. Treasury.

[c]Shown in percent.

[d]1967 = 100.

care's net contribution to per capita health care expenditures for the aged turns out to have been something less than 40 percent in fiscal 1976. If the private health insurance coverage enjoyed by the aged were purchased by them at premiums equal to or greater than the actuarial costs of that coverage, Medicare enrollees themselves may, on average, have paid out of pocket as much as 37 percent of their total outlays on personal health care.

The impact of direct payments on the budgets of the aged is more readily apparent in Table 3. Direct payments borne by Medicare patients are now almost twice as large as they were at the onset of Medicare, a trend often remarked upon in the literature. In terms of constant dollars, however, these payments have remained almost unchanged since the inception of the program, and they have decreased markedly as a percentage of Social Security cash benefits. The percentages would be even lower if they were based on average *total income*, including earnings on assets and from other sources.

One should not draw comfort too quickly from the data in Table 3. First, it has already been noted that premiums for Part B of Medicare and for private health insurance should be added to "direct payments" to obtain the total average cash outlay by the aged for health care. For fiscal 1976, this total would be $541 or roughly 22 percent of Social Security cash benefits in that year. Second, the data cited in Table 3 are per capita figures, averaged over all Medicare enrollees. Because not all Medicare enrollees receive health care in any given year, the average out-of-pocket outlay of those actually sick may be substantially higher than the averages cited in Table 3. Where such persons rely solely or mainly on Social Security as a source of income, their direct payments for health care may create severe budget squeezes.

TABLE 2—PER CAPITA PERSONAL HEALTH CARE EXPENDITURES FOR PERSONS AGED 65 AND OVER, BY SOURCE OF FUNDS (Fiscal Year 1976)

Type of Care	Total	Medicare	Medicaid	Other Public Programs	Direct Payment, Private Insurance, and Other
All Care:					
Amount	$1,521	$653	$244	$134	$491
In Percent	(100%)	(43%)	(16%)	(9%)	(32%)
Hospital Care.					
Amount	689	488	23	116	62
In Percent	(100%)	(71%)	(3%)	(17%)	(9%)
Physicians' Services.					
Amount	256	140	9	2	104
In Percent	(100%)	(55%)	(3%)	(1%)	(41%)
All Other					
Amount	577	24	212	16	326
In Percent	(100%)	(4%)	(37%)	(3%)	(56%)

Source R M Gibson, M. S. Mueller and C. R. Fisher (1977), Chart 2, p. 12.

III. Current Issues in the Medicare Program

Although, by American standards, the Medicare program ranks as a major breakthrough in social policy, it is not much to boast about by international standards. The problem is not only that its administration is cumbersome and that it provides the aged with only incomplete health insurance coverage. Equally troublesome is the fact that the fiscal flow mobilized by the program is sufficiently large to generate demand-pull inflation in the health care sector, yet not large enough to afford the administrators of the

TABLE 3—ESTIMATED DIRECT PAYMENTS FOR PERSONAL HEALTH CARE BY PERSONS AGED 65 AND OVER (Fiscal Years 1966–75)

Fiscal Year	Total Per Capita Expenditures from all Sources	Total Direct Payment in Current Dollars	Average Monthly *OASDHI* Cash Benefit in Current Dollars[a]	Total Direct Payment as Percent of Cash Benefit[b]	Total Direct Payment in Fiscal 1967 Dollars[c]
1966	$445	$237	$83.92	24.2	$244
1967	535	198	–	–	198
1968	647	178	–	–	169
1969	735	206	–	–	190
1970	828	270	–	–	236
1971	925	316	118.10	22.3	262
1972	1,034	367	132 17	23.1	293
1973	1,081	357	162.35	18.3	272
1974	1,181	392	166.40	19.6	275
1975	1,360	390	188.20	17.3	249
1976	1,521	404	207.18	16.2	240

Sources Expenditure series M Gornick, Table 20, p. 18 and Gibson, Mueller, and Fisher (1977), Table 5, p. 9. Consumer Price Index and average *OASDHI* cash benefits: *Social Security Bulletin*, June 1978, Table M-45, p. 69, and Table M-13, p. 43.

[a]As of December of the preceding calendar year.

[b]Shown in percent.

[c]Consumer Price Index for fiscal years calculated from calendar-year series as $C_t = [CPI_{t-1} + CPI_t]/[CPI_{1966} + CPI_{1967}]$ where CPI_t is the reported Consumer Price Index for calendar year t, and C_t is the corresponding index for fiscal year t, with fiscal year 1967 = 100.

program any effective control over the production and pricing of health care services.

How Medicare would fare under national health insurance (*NHI*) is an open question. Many of the *NHI* proposals put forth in the past would have allowed the program to continue as a distinct entity. The continued operation of Medicare seems a virtual certainty under any of the phased-in *NHI* programs currently being considered by the Administration, all the more so because it is one of those benefits-in-kind programs whose immediate economic impact is to provide monetary returns to otherwise unemployed resources owned by the middle- and upper-income classes.

Table 4 indicates the sources of funds for both parts of the Medicare program. While Part A of the program continues to rely almost exclusively on payroll taxes, an increasing proportion of Part B is now being financed by general revenues. This increasing dependence on general revenues results from the linking of *SMI* premiums to Social Security cash benefits which, in turn, have increased less rapidly than medical care prices and expenditures. Whether Congress will remain content to have the program draw ever more heavily on general tax revenues is, of course, an open question.

Currently scheduled tax rates are not expected to maintain the Hospital Insurance Trust Fund (Part A). The trustees expect the fund to be completely exhausted by 1990 Total program costs are projected to average 3.86 percent of taxable payroll during the period 1978–2002. Total scheduled tax rates for Medicare, on the other hand, will be only 2.74 percent of taxable payroll. The resulting deficit for the period is projected at an average of 1.12 percent of taxable payroll.

Because 1990 is far off, however, and

TABLE 4—RECEIPTS OF THE HOSPITAL INSURANCE AND SUPPLEMENTARY MEDICAL INSURANCE TRUST FUNDS, BY SOURCF (Fiscal Years 1967–80)

Fiscal Year	Total Receipts (Billions)	Percent of Total Receipts From			
		Payroll Taxes[a]	Premiums[b]	General Revenue	Interest
Medicare as a Whole					
1967	$4,373	61.8	14.8	22.0	1 4
1970	7,489	64 7	12.5	20.8	2 0
1974	15,419	69.4	11.0	16.4	3 2
1977	22,757	60.0	9.6	26.4	4.0
1980 (Projected)	36,139	64.6	7.9	24.1	3 3
Hospital Insurance Trust Fund:					
1967	3,089	87.5	–	11.0	1.5
1970	5,613	86.3	–	11.1	2 5
1974	11,610	92.2	–	4.3	3 5
1977	15,374	88.9	–	6.1	5.0
1980 (Projected)	25,117	92.9	–	3.5	3.6
Supplementary Medical Insurance Trust Fund:					
1967	1,284	–	50.4	48.5	1.1
1970	1,876	–	49.9	49.5	6
1974	3,809	–	44 7	53.3	2.0
1977	7,383	–	29.7	68.4	1 9
1980 (Projected)	11,022	–	25.9	71.3	2.8

Source: For years 1967–77; *Social Security Bulletin*, June, 1978, Tables M-7 and M-8; for 1980: *1978 Annual Report of the Board of Trustees of the Federal Hospital Insurance Trust Fund*, Table 5, p. 13, and *1978 Annual Report of the Board of Trustees of the Federal Supplementary Medical Insurance Trust Fund*, Table 5.

[a]Includes transfers from railroad retirement fund for Part A.

[b]Some 18,000 individuals not eligible for Part A purchase Part A coverage. These premiums are included in payroll taxes. In 1978, the monthly premium was $54. For Part B, the premiums include those paid by Medicaid on behalf of the aged poor.

because the secular change in the financing of the *SMI* Trust Fund is not necessarily undesirable, neither trust fund per se need be source of immediate concern. At the same time, there is profound concern that the ever increasing outlays on the Medicare program may fail to yield significant medical benefits—at least at the margin—and that the program indirectly displaces outlays for social programs whose benefits are more readily obvious. A number of efforts are therefore underway to constrain the secular growth of outlays per Medicare enrollee, through more vigorous utilization reviews, and through changes in the reimbursement system.

Many economists have argued that the only really effective constraint on the secular growth of health care costs and expenditures will be to let patients share a significant part of the charges generated by their treatment. Since the aged already make substantial direct payments for their health care, however, it is not clear whether much additional cost containment mileage can be had through cost sharing.

Several sleepers may undermine current attempts to constrain outlays for Medicare. First, as the case of end-stage renal disease has illustrated, it will be very difficult for lawmakers to exclude from Medicare coverage other categories of diseases for which relief or a cure is technically feasible, although at enormous costs. Exclusion of coverage for treatable disease might be interpreted by its victims as a death sentence.

Second, it can be expected that the Medicare system will increasingly come to be looked upon as support for the general long-term "social care" of the aged—as distinct from "health care" proper. Although the health care industry is not ideally suited to provide the type of social support needed by the aged, the medical model may be the only vehicle capable of attracting political support for this purpose.

Finally, during the last decade or so health manpower policy in the United States has worked to supply the nation with a generous and possibly excessive supply of health personnel. This supply of manpower constitutes a *demand for health care incomes.* Unlike many other demanders of labor income, the demanders of health care incomes seem to have a remarkable capacity to realize their income aspirations, mainly by virtue of their technical authority on medical matters. Attempts to constrain health care expenditures (health care incomes) in the face of an emerging health manpower surplus may thus turn out to be a frustrating exercise. The aged may, in the end, receive more intensive health care than they need, want, and wish to pay or have paid for.

REFERENCES

R. M. Gibson, M. S. Mueller, and C. R. Fisher, (1977a) "Age Differences in Health Care Spending, Fiscal Year 1976," *Soc. Sec. Bull.*, Aug. 1977, 3–14.

——, ——, **and** ——, (1976) "National Health Expenditures, Fiscal Years 1976," *Soc. Sec. Bull.*, Apr. 1977, 3–22.

M. Gornick, "Medicare Patients: Geographic Differences in Hospital Discharge Rates and Multiple Stays," *Soc. Sec. Bull.*, June 1977, 22–41.

——, "Ten Years of Medicare: Impact on the Covered Population," *Soc. Sec. Bull.*, July 1977, 3–21.

Social Security Administration, *Soc. Sec. Bull.*, June 1978, Tables M-2, M-45, M-46, M-7, M-8, M-13.

——, *Social Security Bulletin Statistical Supplement, 1975*, Washington.

——, *Health Insurance for the Aged and Disabled: Amounts Reimbursed by State and County, 1976*, Washington, prepublication copy.

——, *1978 Annual Report of the Board of Trustees of the Federal Hospital Insurance Trust Fund*, Washington 1978.

——, *1978 Annual Report of the Board of Trustees of the Federal Supplementary Medical Insurance Trust Fund*, Washington 1978.

Social Security Financing and Retirement Behavior

By ANTHONY J. PELLECHIO*

The Social Security retirement program (*OASI*) is the major source of income support for retired workers and their dependents. In 1977 *OASI* paid benefits of approximately $70 billion. Policymakers and the public are justly concerned about its future financial condition. The latest actuarial study by the Social Security Administration (see Francisco Bayo, William Ritchie, and Joseph Faber) concludes that the *OASI* program is not in close actuarial balance over the next seventy-five years. As pointed out in that study, future income and expenditures for *OASI* will depend on labor force participation and the prevalence of retirement, among other factors. The method of financing *OASI* requires that retirees' benefits in each year be paid with contributions (payroll taxes) collected each year as a percentage of workers' earnings covered by *OASI* (the taxable payroll). The theme of this paper is that the size and financing of *OASI* depend on how it affects individuals' retirement decisions.

The basic premise of *OASI* is to pay benefits to individuals whose retirement had decreased their income. But the retirement of individuals who otherwise face the same economic opportunities in terms of their market wage, assets, and other resources can vary due to differences in their retirement benefits. Also, the benefit reduction for earning above the retirement test's exempt amount can change how much people work while receiving benefits. Therefore, rather than being unrelated to Social Security, retirement may be induced by it. If this is so, the taxable payroll is decreased as people shorten and lower their earnings streams later in life. At the same time the lengthening of the period for payment of retirement benefits raises *OASI* expenditures. Thus in the current-cost financial balance retirement is the pivot that determines expenditures as a percentage of taxable payroll. Another way to view this point is that retirement determines the size of the flow from workers' earnings to retirees' benefits. The effect of the Social Security benefit structure on retirement is therefore an important issue in *OASI* financing. The following discussion will examine this issue based on my recent econometric study of retirement behavior.

*Research associate, National Bureau of Economic Research.

I. Social Security as Wealth

The approach here originates with the viewpoint that Social Security is better analyzed on an individual lifetime basis rather than as a transfer system between workers and retirees. This permits identifying Social Security's effects on how much a person works in his life cycle plan for economic activity. Retirement will be emphasized as the relevant margin for observing Social Security's influence on lifetime work. The life cycle approach naturally treats an individual's future Social Security benefits as a form of wealth, a concept presented by Martin Feldstein. A result from my effort at constructing a model for studying retirement is that an individual's Social Security wealth (*SSW*) equals the actuarial present value of future benefits discounted at the market interest rate. *SSW* embodies Social Security's objective of reallocating a person's income to later years to provide support for his own retirement.

A person's *SSW* does not necessarily equal the present value (computed at the market interest rate) of payroll taxes that he pays into the system. If *SSW* is greater than the present value of payroll taxes, then a person's lifetime income is increased by Social Security. Of course if the opposite occurs, then individuals lose lifetime income. This possibility brings to

mind another objective of Social Security—the redistribution of income from high earners to low earners through its benefit structure Redistribution results from meeting a social standard for adequate retirement income. A. Haeworth Robertson also indicates the importance of the tenuous relationship between taxes and benefits for an individual that follows from emphasizing social adequacy at the expense of individual equity. The main point to be made here is that departures from individual equity are responsible for Social Security's potential effects on retirement.

Any change in lifetime income can affect the amount of labor that a person supplies in the market. If Social Security raises lifetime income the life cycle model would imply some decrease in labor supply. This could occur as a reduction in hours worked per year in all years of a person's working life or a decrease in the number of years of work, that is, early retirement. The labor market may not allow individuals to choose freely their hours of work. Consequently, constrained to work full time individuals may choose early retirement as the feasible response to Social Security's lifetime income effect. The opposite responses can occur when Social Security lowers lifetime income. Nonetheless, looking at this effect in a life cycle model alone cannot explain the timing of retirement.

II. The Retirement Decision

A person decides to work based on opportunities available inside and outside the market. In a static setting this involves comparing the value of a person's time in market and nonmarket activity, that is, the values for his work and leisure which are his wage and shadow price of time, respectively. In the life cycle model these values are influenced by Social Security. This occurs because *SSW* is acquired as follows: 1) calendar quarters in which earnings exceed a certain amount are counted toward entitlement for benefits; 2) the benefit amount is based on an average of earnings. Thus, the value of work in a given period is influenced by the effect that earnings will have on *SSW*.

In the quarter which entitles a person to benefits *SSW* increases from zero to the present value of benefits that he could collect in the future when he is eligible (i.e., old enough) for retirement status. Between entitlement and eligibility *SSW* will increase when earnings raise the base used to compute benefits. However, during this period there may be little gain in *SSW* because increases in the base produce less than proportionate increases in the benefit amount. Therefore, the value of work after entitlement may decline relative to that before entitlement (i.e., after earning the jump from zero to positive *SSW*).

Once a person is eligible to receive benefits at age 62 the value of work can decrease further. Postponing retirement shortens the period for receiving benefits. *SSW* can drop because Social Security does not function like an annuity, where benefits foregone in one year raise future annual benefits in order to preserve the capital value of benefit payments. Also benefit payments will be reduced if a person would have worked enough to generate earnings above the retirement test's exempt amount. For these reasons the value of work can decline thereby lowering a person's net wage later in life. This implies that Social Security can have a substitution effect that reduces the number of years of work. This holds for people who either gain or lose lifetime income.

Both the lifetime income and substitution effects induce individuals gaining lifetime income to retire early. These people make the most of Social Security after becoming entitled to *SSW* by collecting it as soon as possible. For those who lose lifetime income the income and substitution effects work in opposite directions. Their retirement is less likely to be affected by Social Security.

III. Empirical Evidence

My empirical research yields significant evidence that Social Security influences retirement. A rich file of data from the Social Security Administration (see F. Aziz, B. Kilss, and F. Scheuren) permits precise calculation of *SSW*. The sample consists of

married men aged 60–70 years entitled to *OASI* benefits, not covered by the railroad retirement system, and not employed by the federal or state government. Also these people did not receive welfare income, unemployment compensation, or disability payments. In this way Social Security's effect on retirement is isolated. A person is said to retire when he stops working before reaching his next year of age in the calendar year 1972. (Defining retirement as working less than half the year or not at all did not change the results.)

The probability of retirement was estimated using probit analysis based on a traditional labor force participation model to which *SSW* was added. The model included market wages imputed from a wage regression that included measures of past earnings. These lagged earnings variables were good predictors of market wage. Other variables used in the model were capital income, schooling, binary variables for rural residence, race and age; wife's wage, schooling, and age. The model was estimated separately for three age groups: 60–61, 62–64, and 65–70.

An increase in *SSW* from $35,000 to $55,000 raised the probability of retirement by .22 relative to a .78 retirement rate in the 65–70 age group. This *SSW* increase is approximately a two-standard deviation change centered on the mean value. For 62–64-year olds this increase raised the probability by .15 relative to a .41 rate. These probability increases apply to $4–7 hourly market wages and decline outside this range. This indicates that individuals retire when *SSW* is high, holding market wage constant. This implies that people collect *SSW* when it is likely to exceed the present value of payroll taxes they pay. Thus, *OASI* expenditures are increased because it does not function like an annuity.

A quasi experiment with a control group is performed by estimating the model over a sample of individuals ineligible for retirement benefits. Since 60–61-year olds are entitled to but cannot receive benefits, *SSW* should not affect their retirement decision. The estimated response to *SSW* was small and insignificant.[1] This supports the conclusion that the observed effect on eligible persons is a causal relationship. At least some of the arguments that the *SSW* effect is spurious due to left-out variables or a non-linear or more complicated specification are addressed by this quasi experiment. It also suggests that retirement may be the relevant margin for altering lifetime work in response to Social Security.

Some brief remarks on the retirement test are appropriate here. Robert Ball points out that the retirement test is undoubtedly the most unpopular Social Security provision. Both Larry Kotlikoff and I have constructed earnings distributions for individuals subject to this test in each year from 1967 through 1974, a period in which the exempt amount assumed four different values. There is a striking concentration of people just below the exempt amount in all years (see Kotlikoff) Also, Ball points out that in 1975 nearly one-fourth of the beneficiaries with earnings below the exempt amount of $2540 were concentrated in the $2100–2500 interval. That individuals change their behavior in response to Social Security seems evident.

Some qualifying remarks are required at this point. The above results show that people work less later in life but they may work more in earlier years in anticipation of early retirement. Also, if Social Security can increase lifetime income the value of work is raised at least for earning entitlement. This encourages work before retirement. Identifying and estimating lifetime income and substitution effects on labor force participation separately in both early and late periods of the life cycle is yet to be done.

IV. Conclusion and Framework for Reform

This discussion proposed that Social Security can and does induce retirement. The main inference to be drawn is that benefit payments without work disincentives can

[1]The probit estimate for *SSW's* coefficient for 60–61-year olds was one-fifth and one-eighth that for 62–64- and 65–70-year olds, respectively. Furthermore, it was one-half its standard error in magnitude.

improve *OASI* financing. Possible reform of Social Security follows from looking at its two objectives discussed here which are: 1) reallocation of a person's lifetime income to provide support for his own retirement; 2) redistribution of income between individuals to provide socially adequate support.

Reallocation of lifetime income can be accomplished without income or substitution effects if a person's future benefits equal payroll tax contributions in present value terms at a fair market and actuarial return. In other words, Social Security should function as an annuity and the retirement test should be eliminated. If a person decides to collect an annual retirement benefit of $5,000 at age 62 or $25,000 at age 75, *OASI* neither gains nor loses revenue on his individual account. As an annuity that achieves individual equity the system will not distort retirement decisions.

General revenue can finance the second objective because the income tax system is the proper place for redistribution. Poverty among the aged should not be addressed in a retirement pension program. A single program of income maintenance can provide socially adequate support based on lifetime or permanent income. In this way poverty is not misconstrued as a problem of the aged. It is important to emphasize that general revenue covers the cost of redistribution now done by Social Security—this must be carefully calculated and not done in an arbitrary way. Also, a general income maintenance program can have work disincentives of its own.

At present through a complicated benefit structure Social Security tries to meet both objectives at once which confuses the issues and causes problems. Providing generous support to low earners induces retirement and raises *OASI* expenditures. High earners are implicitly paying for redistribution under the guise of contributing to their own retirement. The size and financing of *OASI* depend to a considerable degree on how much retirement Social Security, as it presently operates, will make people choose. As the age distribution of the population shifts the continued labor force participation of older workers will be important to the economy. There should be no disincentives for their participation.

REFERENCES

F. Aziz, B. Kilss, and F. Scheuren, "1973 *Current Population Survey*—Administrative Record Exact Match File Codebook," studies from Interagency Data Linkages, rept. no. 8, Social Security Administration, Office of Research and Statistics, 1978.

Robert M. Ball, *Social Security*, New York 1978.

F. R. Bayo, W. D. Ritchie and J. F. Faber, "Long-Range Cost Estimates for Old-Age, Survivors and Disability Insurance System, 1978," actuarial study No. 78, Social Security Administration, Office of the Actuary, Washington.

M. Feldstein, "Social Security and Saving: The Extended Life Cycle Theory," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 77–86.

L. J. Kotlikoff, "Social Security or Social Insecurity? The Choice Is Ours," unpublished paper, Univ. California-Los Angeles, June 1978.

A. J. Pellechio, "Social Security and Retirement Behavior," unpublished doctoral dissertation, Harvard Univ. 1978.

A. H. Robertson, "The Outlook for Social Security," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 272–74.

Alternative Economic Policies for the Revitalization of U.S. Central Cities

By CLEVELAND A. CHANDLER AND WILFRED L. DAVID*

The formulation of suitable policies and strategies for improving the economic viability of *U.S.* central cities constitutes one of the most crucial tasks facing policymakers in the urban field. The problem has become increasingly severe for the large number of central cities in the North East, North Central, and Western states. The basic issue is one of determining the character, dimensions, and magnitudes of policies that will reverse the process of cumulative decay and propel urban core areas into a new orbit of self-sustaining development.

This paper examines some structural dimensions relating to change in central cities, and how these can be influenced by means of concerted policy action. An underlying assumption is that most policies for urban change tend to produce limited results because they are anchored on unsuitable conceptual underpinnings. The paper therefore attempts to develop a systematic framework of analysis with a view to highlighting some of the major structural dimensions of the urban environment which are amenable to those kinds of policy interventions necessary for revitalization.

The paper is divided into three parts. Section I briefly outlines the conceptual focus. This is illustrated by a simple policy model in Section II. The policy implications for central city revitalization are drawn in the final section.

I. Conceptual Underpinnings

The central city can be considered the core of a metropolitan area or of other functional economic regions encompassing one or more states that make up the nation. In recent times questions have been raised pertaining to what consitutes the proper unit of urban analysis—the central city, Standard Metropolitan Statistical Areas (*SMSA*s) or what Brian Berry calls "daily urban systems." The legitimacy of such a concern notwithstanding, this paper takes the more convenient route of assuming that the central city is a functional area of analysis.[1] As such, the main concern is with the endogenous factors conditioning the economic viability of such cities as well as the impacts of exogenous policy actions emanating from national, state, regional, or metropolitan sources.

From a policy perspective, the typical central city can be visualized as a complex political socioeconomic system depicted by the continuous interplay of a large number of variables and the interdependent processes which govern their behavior. This suggests a systems approach as the correct frame of reference. While there are several examples of this type of study of the urban environment (see for example, Jay Forrester), in very few cases has emphasis been given to study and unravelling the dynamic cumulation of forces responsible for central city growth, decline, or

*Professor and chairman, department of economics, and professor of African studies, Howard University, respectively. We share equal responsibility for writing the paper. We wish to thank Willie Taylor for useful comments; Ravi Aulakh, A. Khanna, Jane Li for research assistance; and Pam Neverson for typing. Remaining errors are our responsibility.

[1]For practical purposes central cities are defined here as political jurisdictions of *SMSA*s which contain the main central business district, the location of the local government, the residential community, as well as satellite commercial, shopping, and industrial districts within central city limits. All components of a central city are considered a unit.

stability. And in general the study of the problems facing central cities and urban areas has relied on the neoclassical framework of analysis, with a primary focus on equilibrium conditions rather than on the disequilibrium processes which characterize urban environments (see Harry Richardson).

The dynamics of any socioeconomic system can be explained by the fact that there is usually circular causation among its endogenous conditions, that is, if one condition changes, others will change in response. In general, such secondary changes will induce new changes in all other parts of the system, in some cases even reaching back to the initial change. Such a situation of general interdependence, with everything causing everything else, means that as the process cumulates, there may be no unique equilibrium. The system can move in the direction either of cumulative growth, cumulative decline, or it may continue to turn around on its axis (see for example, Gunnar Myrdal, 1944, 1957; Albert Hirschman).

Here a distinction should be drawn between the primary (initiating) changes and the induced responses. In the majority of socioeconomic systems the primary changes and the induced responses tend to move in the same direction and have cumulative effects. As the process develops, more changes are induced through a system of feedbacks, with a resulting disproportionality between the primary change and the cumulative growth or decline of the system which ensues.

The primary focus of this paper is on those initiating impulses that take the form of exogenous policy changes (for example, increased federal investment) which can be used to change one or more endogenous conditions in the central city system. In many cases, however, the scope and direction of such policy intervention normally constitute reactions to changes taking place among the endogenous factors. Due to space constraints, we were forced to present above what is at best a very incomplete picture of the general nature of the cumulative disequilibrium process. The detailed ramifications are by now well-known to members of the profession. However, to date there have been relatively few attempts to apply this analytical framework to the study of urban policy questions. (See William Baumol 1967, 1972; Wallace Oates, E. Philip Howrey and Baumol; Peter Albin.) The policy model which we have developed in the next section follows in this tradition.

II. A Model of Central City Degeneration/Revitalization

The purpose of the following model is to demonstrate how the problem facing the majority of central cities can be conceptualized in terms of a disequilibrium process. The five equations listed below do not fully represent a complete set of functional relationships characterizing the central city environment. Further, the equations are written in linear form for purposes of exposition and manipulation. A model written for purposes of econometric testing can be more fully specified and problems relating to multicollinearity, nonlinearity, etc., more fully addressed.

$$(1)\quad Y_{t+1} = \alpha_1 E_{t+1} + \alpha_2 N_{t+1} + \alpha_3 M_{t+1}$$

$$(2)\quad E_{t+1} = \beta_1 I_{t+1} + \beta_2 G_{t+1} + \beta_3 H_{t+1}$$

$$(3)\quad M_{t+1} = \psi_1 I_t + \psi_2 G_t + \psi_3 H_t$$

$$(4)\quad I_{t+1} = \delta_1 G_{t+1} + \delta_2 Y_{t+1}$$

$$(5)\quad H_{t+1} = \rho_1 G_{t+1} + \rho_2 Y_{t+1}$$

where Y = average per capita income level as a proxy for overall economic well-being; E = average employment rate; N = population as a proxy for city size; M = percentage of population above poverty level (middle and high income); I = private investment in nonfinancial assets; H = new housing investment or remodelling expenses; and G = government expenditure.

The above equations are based on well-established ideas about central city structure and dynamics. Equation (1) postulates that income levels (overall well-being) depend on the average employment rate, the proportion of central city residents earning well above the poverty levels, and city size. City size enters the equation as an exogenous variable, but nevertheless poses a constraint on the

process of urban development because of the positive relationship with agglomeration economies. The process of revitalization requires the creation of conditions which would bring about self-sustained economic development. An important determining factor here is the employment process which in turn depends on the industrial structure, the nature of human capital formation and related factors which are well-known. An attempt has been made to capture such dependencies in equation (2) in which the average employment rate is made to depend on private investment and government expenditure, as well as the rate of buildup of the housing stock in both quantitative and qualitative terms. Other important variables such as nonhousing consumption factors and net export demand are excluded.

The proportion of the central city population earning adequate incomes (equation (3)) depends on previous levels of private investment, government expenditure, and housing investment. These factors tend to create and sustain those kinds of environmental conditions which maintain incomes at a high level and reduce urban blight and the flight to the suburbs.

Private industrial investment (equation (4)) depends on government expenditure and income growth. A similar process obtains in the case of private investment in housing (equation (5)). An underlying hypothesis here is that the nature of the urban development process depends on the interdependent relationship between public and private sector investment. It is known that an important share of capital stock of central cities consists of physical and social overhead capital of various forms which in itself is predetermined by the spatial distribution of people and economic activities. The scale and distribution of this public capital usually generates changes in private investment. The process works as follows: an exogenous injection of different mixes of government expenditure increases the size and quality of the public capital stock which in turn induces private investment. The increase in private investment changes the location decisions of households and business firms, and once the system takes off in this manner the subsequent growth process becomes self-sustained and cumulative.

Thus, assuming that the initiation of the development process is not constrained by city size, the strength of the revitalization process (i.e., self-sustained economic development) would depend on the exogenous injection of public resources. In designing a revitalization policy, careful attention has to be paid to several questions. Of crucial importance is the form which the injection of public capital takes; that is, whether it is made merely in response (passive) to changes in population as distinct from more active types which generate expansion by inducing private investment. An active policy for revitalization should be based on an optimal mix of public investment between "people" vs. "places" (see Anthony Downs) and address problems of the durability of such investment, its sequential and scale characteristics, etc. (see Richardson). An important question also relates to the manner in which such public injection of capital is financed, and the implications for such financing would have to be carefully considered.

After manipulation of equations (1)–(5) we obtain a solution which may be written as

$$Y_e = \frac{(l_1 + l_o)G}{k_o - k_1} \tag{6}$$

which reveals several points of interest to the policymaker interested in revitalization strategies. While the convergence hypothesis has been verified for certain types of regions (see for example, George Borts and Jerome Stein), in the majority of cases the process is a discontinuous one. The presence of dynamic disequilibrium forces mentioned by several authors leads us to believe that there is no automatic tendency towards equilibrium and nonconvergence is the rule rather than the exception. Second, whether policies and strategies for revitalizing central cities lead to cumulative expansion or degeneration would depend on the size and behavior of the growth multiplier.

One possible way of defining this problem is

in terms of maximizing the sum of incomes of all central cities providing that the growth of such incomes for the poor and declining central cities is at least equal to the mean rate of income growth of all cities. To illustrate, supposing that it is possible to identify, say, four types of central cities with per capita income growth rates c_1, c_2, c_3, c_4, the weights of which in terms of income being a_1, a_2, a_3, a_4, with the richest class of central city a_1 having the fastest rate of growth c_1, and so on in descending order of performance. Assuming that the fourth class of central city is the poorest and suffering from degeneration, its rate of growth would be inferior to the average national rate C. The policy objective therefore becomes one of making the growth rate of this class of central city at least equal to the national average, that is, $c_4 \geq C$ with $C = a_1 c_1 + a_2 c_2 + a_3 c_3 + (a_4 - 1)c_4 \leq 0$.

Under this policy the target level of per capita income (Y_e) shown in equation (6) and its rate of growth should not therefore be inferior to the national average levels and rate of growth. The achievement of such a target for lagging central cities depends (again referring to equation (6)) on the nature of the exogenous injections of public spending, the endogenous conditions shaping the central city environment as reflected by the coefficients l_o and l_1, as well as the behavior of the growth multiplier which is constrained by k_o and k_1. Some broad policy implications are drawn below.

III. Policy Implications

So far, the analysis clearly indicates a multidimensional and comprehensive approach with multiple policy objectives. While the initial target is some adequate level and rate of growth of average per capita income related to some national norm or mean, the structure of equations (1)–(5) clearly implies the multiple objectives of adequate household income, full employment, the eradication of poverty, diversification of sources of household incomes, private investment in the local housing and business sectors by indigenous households, and increased public expenditure with tax reform.

Policymakers are faced with a basic policy choice between a passive adjustment approach toward stagnating equilibrium decline, that is, the covergence hypothesis; or the cumulative disequilibrium option toward revitalization of central cities from the inside out rather than from the outside in. The appropriate policy instruments for the latter approach are public expenditures and investments at the federal, state, and local levels that will induce sustained desired responses in the core endogenous conditions of central cities. Clearly, some critical minimum volume and composition of public investment is required to initiate the process of structural transformation and institutional change.[2]

The required level and mix of public expenditures is injected through policy variables such as household income from all sources, private saving and investment from the local household sector to the business sector, and improvement in stocks of human and nonhuman capital. Both the public expenditure and the policy variables would be calibrated to impact most directly on target populations of individuals, households, business firms, institutions, and community organizations.

While the ends and means for economic revitalization are directed toward endogenous conditions of distress and potential in central cities, this does not imply isolation or exclusion from suburban, rural, and other parts of a metropolitan region. Desirable indirect spillovers are too numerous to cite here. However, all costs and benefits to subdivisions would be considered in the context of comprehensive development planning for the entire metropolitan region. Interdependence is emphasized as an alternative to dependence or independence among components of the region.

Although the initial injection for activating

[2]Initial injection about $10 billion of new federal dollars for central cities or about 1 percent of *GNP* and 3–5 percent of the federal budget is proposed. Fractional amounts of state and local investments would accompany the federal expenditure and activate the multiplier-accelerator mechanism of the central city private economy. The proposed initial federal injection is about twice the amount of new federal money in the recent National Urban Policy Program of the Carter Administration.

the process of revitalization of central cities is external to their present capacities, it would be restricted to the funding of capacity building programs. Thus, households, business firms, and institutions in central cities are expected to sustain initial departures from decline and stagnation to structural transformation and revitalization.

REFERENCES

P. S. Albin, "Unbalanced Growth and Intensification of the Urban Crisis," *Urban Stud.*, June 1971, *8*, 139–46.

W. J. Baumol, "Macroeconomics of Unbalanced Growth," *Amer. Econ. Rev.*, June 1967, *57*, 415–26.

Brian J. L. Berry, *Growth Centers in the American Urban System*, Cambridge, Mass. 1972.

George H. Borts and Jerome L. Stein, *Economic Growth in a Free Market*, New York 1964.

A. Downs, "Urban Policy," in Joseph A. Pechman, ed., *Setting National Priorities: The 1979 Budget*, Washington 1978.

Jay W. Forrester, *Urban Dynamics*, Cambridge, Mass. 1969.

Albert O. Hirschman, *The Strategy of Economic Development*, New Haven 1958.

Gunnar Myrdal, *An American Dilemma: The Negro Problem and Modern Democracy*, New York 1944.

——, *Economic Theory and Underdeveloped Regions*, London 1957.

W. E. Oates, E. P. Howrey, and W. J. Baumol, "The Analysis of Public Policy in Dynamic Urban Models," *J. Polit. Econ.*, Jan./Feb 1971, *79*, 142–53.

M. Peston, "The Dynamics of Urban Problems and its Policy Implications," in Maurice Peston and Bernard Corry, eds., *Essays in Honor of Lord Robbins*, London 1972.

Harry W. Richardson, *Regional Growth Theory*, New York 1973.

Hospital Production—Can Costs be Contained?

By CHARLES E. ANDERSON*

Although the modern hospital complex has been heralded as a center of medical care advancement and cure, its market performance has been faulted on several grounds. The criticisms which advocate hospital reform and structural reorganization at the industry level address problems of rising hospital costs, productive inefficiencies, and underutilized resources.

Within the past several years, many states have acquired legal powers to regulate hospital costs through prospective reimbursement formulas as well as the monitoring of both capital expenditures and new operating expenses. Prospective reimbursement is defined as the "financial reimbursement of health care providers where the amount or rate to be paid is established prior to the period over which the amount or rate is to be applied "(see Frank Sattler, p. 3). The monitoring of capital expenditures and new operating expenses is commonly referred to as "certificate of need review procedures" (see Clark Havighurst, p. 2).

Whether a judicious use of these regulatory measures will be sufficient to effect a quasi-competitive structure in the hospital industry is a question posed by the majority of hospital planners and policymakers. Generally it is felt that the broad thrust of cost containment policies will improve the allocation and productive efficiency of community health resources and that, to a large extent, these goals may be achieved without sacrificing the quality and quantity of hospital output and without violating preestablished cost boundries. However, the outcomes of these procedures seem uncertain. First, there is no empirical basis for such positive expectations. Indeed no state plans for prospective reimbursement and/or certificate of need procedures have been fully implemented. Secondly, and most pertinently, current optimisms rests on strong assumptions pertaining to the behavior and authority of hospital administrators. (See Sattler and Katherine Bauer.)

This paper contributes to current discussion along two lines. Essentially, it will discuss the cost and output effects of two types of regulatory structures imposed on a hospital industry which is characterized by large firms and limited price and quality competition. The regulatory structures, although distinct, differ only in degree; they represent the end points of a class of controls often proposed for the hospital sector.

The first system moderates industry cost levels by reimbursing hospitals on the basis of average industry costs and the classification of firms. Consequent production and investment decisions are determined by market forces. The second system moderates industry cost levels along lines specified by need and community preference criteria. In this arrangement, oligopolistic firms deliver medical care in a system largely shaped or prescribed by the joint decisions of the medical care profession and planning agencies rather than wholly by market forces. The policy instruments at the disposal of planning agencies in this framework are rate setting and certificate of need review procedures. Thus, investment outlays are jointly determined by planning agencies and hospital administrators; only price and quantity decisions are viewed as the sole responsibilities of the administrators.

The first type of regulation will be referred to as a limited competition hospital industry. The second type will be referred to as a fully planned hospital industry.

I. A Limited Competition Hospital Industry

The imposition of prospective rates will increase cost efficiencies among hospitals if hospital administrators are influenced by

*Assistant professor of economics, Livingston College, Rutgers University. I am thankful to Roger Mack of SRI International for suggestions and comments which substantially improved this paper.

financial considerations and if the income of hospitals can be strongly related to operations. Under these assumptions financial incentives will, in part, determine administrative decisions concerning hospital operations (see Sattler). An ideal prospective rate schedule, given a market stance, will strengthen the competitive advantage of relatively efficient firms and, conversely, place relatively inefficient firms at a competitive disadvantage (see Paul Feldstein).

Consider Figure 1. The schedule *NPC* represents the expected demand schedule, net of insurance, for hospital service Q. It relates the maximum price (i.e., net patient charge) that can be levied per unit of output by the hospital. A retrospective reimbursement system grants hospitals lump sum payments for services produced that are at least equal to the difference of production costs and net patient charges (the vertical distance between its average costs, *CF*, and *NPC*). In Figure 1, two hospitals whose cost curves are denoted by *CF1* and *CF2* receive different unit subsidies given that they produce the same quantity of services. Hence, inefficient hospitals have no incentive to exit from the market or adopt more efficient modes of production or organization.

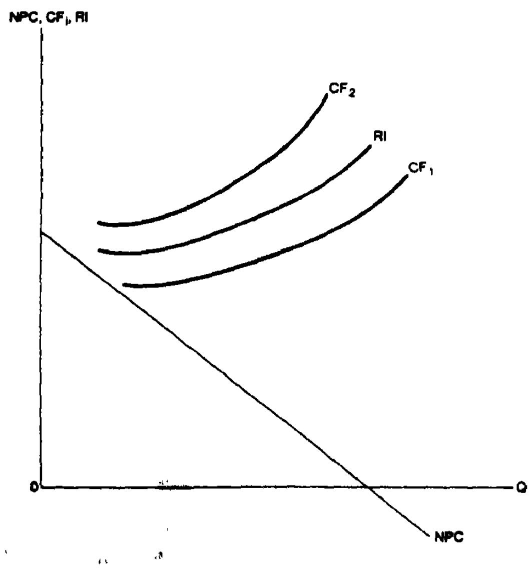


FIGURE 1

In contrast, prospective reimbursement systems of the type discussed here induce efficiency incentives by limiting hospital cost compensation to an amount equivalent to average industry costs given appropriate allowances for hospital classification. In Figure 1, both hospitals would receive a unit subsidy amount equal to $RI - NPC$ at equal output levels. Here *RI* is a compensation schedule based on average industry costs and hospital classification given the production of Q.

To generalize the point, let hospital revenues be denoted by the expression:

$$(1)\quad R = TNPC(Q_1, \ldots, Q_k) + S(Q_1, \ldots, Q_k)$$

where *TNPC* and S denote gross hospital revenues from net patient charges and prospective reimbursement compensations, respectively.

Similarly, let total costs of production be denoted as

$$(2)\quad TC = TC(Q_1, \ldots, Q_k)$$

where Q stands for the range of services produced by the hospital. If hospital administrators are either utility or benefit maximizers, as is often supposed, they maximize a preference function, $U(Q_1, \ldots, Q_k)$, subject to the budget constraint, $\pi_o = TNPC + S - TC$, where π_o denotes a target profit level. First-order conditions of equilibrium are given by the expression

$$(3)\quad \frac{\partial U}{\partial Q} = \lambda\left(\frac{\partial TC}{\partial Q_i} - \frac{\partial R}{\partial Q_i}\right), \quad i = 1, \ldots, k$$

$$(4)\quad \pi_o = R - C$$

The equilibrium conditions of the utility-maximizing firm for two possible cost situations is depicted in Figure 2 for a particular good.

If the hospital firms were profit maximizers, they would produce at the point where $MC' - MR' = 0$, denoted by Q_2 and Q_1, respectively, in Figure 2. Thus the profit maximizer produces less and charges more per unit of output than the utility maximizer.

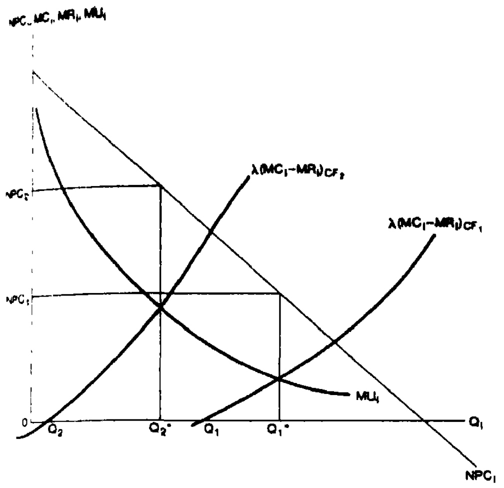


FIGURE 2

Differences in the cost situation of the firm clearly makes a difference in the amount of output produced by the hospital and the unit prices it will charge for the output. Hospitals that are similarly classified, but whose cost schedules differ will produce different levels of output and will charge different prices per unit of the same grade. Hence, the policy dynamics of a reimbursement scheme such as the foregoing are twofold; first, it forces marginal firms from the marketplace; second, it accentuates differences in the cost structures of competitive hospitals and possibly magnifies differences in the price structures of competitive hospitals. Given an appropriate level of information, competitive pressures will favor those hospitals that are more cost and production efficient. The net effect will be lower industry cost.

II. Problems and Prospects of a Limited Competition Hospital Industry

It is doubtful that a rate setting scheme can be employed effectively except within a broad and discretionary policy setting. Indeed the drawbacks of rate imposition, as described here, are both structural and behavioral. Structurally, given that reimbursement constraints are effective at the industry level, the additional requirement that individual hospitals operate efficiently cannot be readily assured; rate schemes do not necessarily promote internal efficiencies within the firm.

Concerns pertaining to hospital management bring into question whether prospective rate schemes can be effective in a limited competition industry. To begin with, even if rate schemes accord hospitals a positive subsidy for being relatively efficient, a large percentage of administrators may expand expenditures to the levels of their reimbursement schedules without significantly increasing output. A sufficient condition for such behavior is that hospital administrators be prestige or quality maximizers. This, of course, implies various types of satisficing behavior. Such satisficing behavior would have the effect of dampening policy action through shifting the long-term reimbursement schedules of the industry continually upward. Eventually, inefficient hospitals could survive through default.

Secondly, many question the undue emphasis that has been placed on the role of the hospital administrator with respect to hospital management and operations. In all likelihood, the administrator has neither the authority nor the control over hospital operations that so often have been ascribed to him. At best, the administrator may affect decisively the composition and range of services offered by the hospital over the long term; and then only if it is conceded that he has the powers to eliminate, promote, or initiate departments and operations of varying levels of promise. The more realistic assertion is that physicians and administrators jointly determine the use and management of the hospital. The decisions of the former may counter the policy intent of rate schemes unless these decisions are commensurate with those of hospital administrators.

III. The Fully Planned Hospital Industry

The social management of a limited competition and a fully planned hospital industry differ critically in several respects. In summary, comprehensive health care proposals for fully planned hospitals advocate that the hospital sector be structured in the follow-

ing manner (see David Pearson and Milton Roemer et al.). First, the organization and physical distribution of medical care facilities, at least initially, should be based on need criteria rather than free determination by market forces. Objectives dictate that comprehensive personal health services be provided and extensively coordinated within industry regions. Further, extensive information and screening systems are to monitor disease incidence and relative population health status by designated region. Second, the growth and expansion of the industry is more often described as planned rather than determined by free market forces. Growth objectives are promoted along political and economic lines. The political structure of regulation allows full community expression of preferences regarding the structure, delivery, and distribution of services. The economic component of the system provides, through exogenous but market-related instruments, the wherewithal for planning agencies to realize cost, investment, and distribution objectives. Third, a reasonable financial structure supports the system. Most proposals suggest financing through tax funds, insurance contributions, or a combination of the two.

Clearly, within such an overall planning framework rate setting and certificate of need policies assume a broader regulatory role than management policies in hospital industries of limited competition. In particular, the uses of industry cost information for policy purposes differs sharply in the social management of the industry types discussed. In the case of a limited competition industry rate management serves to encourage the exit from the industry of marginal firms. Investment decisions and the range and quality of medical services are determined indirectly through rate policy actions and the consequent production decisions of suppliers.

In a fully planned hospital industry rate management policies serve also to effect socially acceptable levels of hospital costs. But, additionally, the market effects of policy actions are further subject to the discretion of planning agencies, given distribution and investment objectives. Essentially, this means that marginal hospitals are evaluated from other vantage points than solely that of relative cost; the decision of a firm to enter and exit from the market is jointly determined by planners and administrators. Second, imitation of superior firms within classifications may be encouraged by planning agencies. Third, industry cost data identify hospitals whose facilities use substantially deviates from normal utilization patterns; for hospitals of designated size categories, aberrant utilization patterns signal that facilities should be either expanded or converted to alternative use.

Likewise, certificate of need procedures are perceived differently as policy instruments by planners in the two types of industries. In both systems, review procedures lend leverage to the cost and investment strategies of planners Conferral, however, is motivated by distinct intents. The rationale for review procedures in systems of limited competition is to discourage excessive capital formation at industry levels. In the planned system, strong emphasis is also placed on community needs or requirements; capital expansion tends to be in accord with community preferences.

In these respects, greater management flexibility is associated with fully planned hospital systems. In the limit, an astute use of prospective rate schemes and need review procedures imply a highly dynamic and efficient hospital industry within a fully planned context. Rate setting and need review procedures can be employed to expedite an industry-wide diffusion of cost-reducing technological and organizational innovations in the area of medical care production, management, and delivery. Second, a planned hospital system promises, relative to limited competition industries, a better articulation and coordination of need and consumer demands. Need is established by health authorities; consumer demand refers to personal medical wants effectively expressed in medical care markets.

IV. Problems and Prospects of a Fully Planned Hospital Industry

Peculiarly, the problems inherent in the concept of a fully planned hospital industry center about those structural attributes

which, many claim, lend such systems rational credence. For example, to allocate medical care facilities according to need criteria is suspect, not only because need criteria poorly anticipate demand levels for medical services, but most likely because they ignore demand considerations altogether. As such, allocation principles based on need do not guarantee adequate levels of capital formation, nor do they affect any tendencies toward equilibrium capital levels in the hospital industry. Second, need criteria are not constant variables (see Michael Cooper). They vary substantially with minor changes in the discretions or preferences of physicians and consumers. Also, health authorities may perceive need criteria as open-ended variables with no limits to be placed on the quality and range of services to be offered by institutions

Certainly, the tenets of joint management advanced by comprehensive health care proposals imply an extensive bureaucracy and, thus, high administrative costs. But a large part of total management costs are bound to be associated with the transactive and agreement processes that regulate interaction among planning agencies and hospital administrators. In particular, costs associated with agreement processes are expected to be highly volatile. In the larger number of instances, concensus among planners and administrators, given equally weighted preferences, will not be routinely attained regarding industry direction and change. Recourse to the legal settlements of disputes will surely be a lengthy as well as costly process. Finally, one expects that prospective rate schemes of the type described here will influence hospital behavior in the same manner described in previous sections. Any difference in degree will be due to the extra intervention of planning agencies.

V. Conclusions

As evaluated here, neither of the systems presented fully satisfy social demands for a hospital industry that moderates hospital costs and, yet, effects an acceptable if not efficient allocation of health care resources. A limited competition industry permits increasing hospital costs in the long term and in spite of reimbursement contraints, if administrators are satisficers or physicians perceive hospital resources as zero priced inputs. Fully planned hospital industries fail in that they virtually negate market forces. The requirement of joint agreement among health authorities and planning agencies regarding industry expansion and control severely limits the effective market decisions of administrators.

Appropriate modifications of either of these systems should allow markets to play a viable role in determining socially acceptable ranges of prices and costs for hospital services. Bestowing project veto powers on planning bodies in the limited competition industry would curtail satisficing behavior and provide incentives for stronger cooperation between physicians and administrators. Deemphasizing need criteria and centralizing planning activities in the fully planned hospital industry return discretionary powers to hospital administrators in investment matters but maintain physician interest in efficient hospital management.

REFERENCES

K. Bauer, "Hospital Rate Setting," in Michael Zubkoff and Ira Raskin, eds., *Hospital Cost Containment*, New York 1978.

M. H. Cooper, "The Economics of Need," in Mark Perlman, ed., *The Economics of Health and Medical Care*, New York 1974.

P. Feldstein, "An Analysis of Reimbursement Plans," HEW-SSA, res. rept. no. 26, 1968.

C. Havighurst, "Regulation in the Health Care System," *Hospitals*, June 16, 1974, *48*, 51–54.

D. A. Pearson, "The Concept of Regionalized Personnal Health Services in The United States," in Ernest W. Saward, ed., *The Regionalization of Personal Health Services*, New York 1976.

Milton Roemer et al., *Planning Urban Health Services*, Boston 1975, ch. 10.

F. L. Sattler "Hospital Prospective Rate Setting," HEW-SSA, final rept., 1974.

Housing Segregation and Black Employment: Another Look at the Ghetto Dispersal Strategy

By SAMUEL L. MYERS, JR. AND KENNETH E. PHILLIPS*

There is some irony in how the debate about whether black inner-city ghettos should be dispersed or developed has been translated into public policy on revitalization of central city areas. Recall how, a decade ago, vehement defenders of ghetto dispersal were stating that black employment opportunities had been restricted because of housing segregation and suburbanization of jobs. Along with low automobile ownership among ghetto residents and declining low-skill job opportunities in the central business district, housing segregation and suburbanization of jobs were factors ranking high on the list of explanations of why gilding of ghettos should not be encouraged. The alternative offered was ghetto dispersal.

Recall, on the other hand, how opponents of ghetto dispersal denied the basic assumptions of its defenders. It was denied that suburban jobs were any better, opportunities any less bleak, or labor markets any less discriminatory for blacks living in segregated ghettos than they were for blacks living in suburban areas. The advocates of developing inner-city communities proposed policies of public employment, subsidies for renovation of commercial and residential property, and relocation of businesses back to the central cities. Many of these policies have been adopted and some have been modestly successful in curtailing continued dehabilitation of our central cities. Yet, the major impact of recent urban revitalization schemes, particularly plans to renovate inner city residential property, has been massive dislocation of poor blacks. Thus, in many cities, the ghetto had indeed dispersed. Who would have guessed ten years ago that the very policies advocated by the proponents of inner-city community development would hasten ghetto dispersal?

This essay revisits the decade-old ghetto development vs. ghetto dispersal debate. Whether by design or by accident, ghetto dispersal may be a suboptimal means of improving the economic status of poor urban blacks. There are a number of theoretical arguments besides those offered a decade ago for why decentralization of black residential location alone need not resolve the problems of low incomes, high unemployment, and lack of economic mobility of inner-city residents. One argument states that the observed wage differences between suburban and ghetto jobs can be attributed to the distance that workers must travel to accept these jobs and not merely to the differences in the labor market demand. Hence moving ghetto workers closer to suburban jobs, thereby reducing their travel costs, may merely reduce the premium required to attract distant workers. Another argument is based on a model of job search in which moving ghetto workers closer to suburban jobs reduces their search costs and hence their reservation wages rise. In the short run, the effect could be longer durations of unemployment. Only in a world of competitive labor markets, no business cycles, and no changes in aggregate demand could we guarantee that in the long run reemployment wages would rise.

The empirical evidence is mixed. In a number of isolated cases wages and employment really are higher in suburban labor markets. Support is found for the notion that the demand for low-skilled workers exceeds the supply in suburban areas. Moreover, poor blacks and whites appear to have similar employment experiences in the suburbs. The isolated evidence does not support the view of racial discrimination in suburban labor markets.

However, support is found for the view that

*The University of Texas-Austin and the Rand Corporation, respectively.

there are positive externalities in employment in one's own community. The isolated evidence suggests that poor workers who both live and work in census tracts of similar racial composition have better employment experiences.

I. The Original Case for Ghetto Dispersal

Can decentralization of residential locations improve black employment prospects? This issue, of course, is but a part of the larger "ghetto development vs. ghetto dispersal" debate. While the dispersal of the urban ghetto has been debated on a number of levels, perhaps the igniting spark, at least among economists, can be traced to John F. Kain.

Kain explored a number of hypotheses relating housing segregation to employment. A reasonably accurate premise is that job opportunities have become spatially dispersed throughout the metropolitan area since World War II. Principally, labor market opportunities have been expanding in suburban areas while blacks have remained segregated in housing centered near the urban core. The resulting spatial separation of jobs and residences, Kain contended, would impose high transportation costs on ghetto workers because of the long distance which those workers travel. The net effect of these imposed travel costs is to reduce the effective wage which central city workers receive relative to white suburban residents. Still another cost imposed by the spatial separation of jobs and residences includes the higher search costs which blacks experience in finding suburban employment and the lower quality of information about potential job opportunities.

Other issues explored by Kain included the possibilities that employers outside of the ghetto discriminate against blacks and those in the city discriminate in favor of blacks. These possibilities are offered as explanations for how residential segregation affects the spatial distribution of employment. Coupled with two additional hypotheses, 1) that residential segregation reduces black job opportunities, and 2) that suburbanization of jobs worsens the relative position of ghetto workers, Kain presents his case for dispersal of the ghetto.

II. The Case Against the Case for Ghetto Dispersal

Kain's hypotheses have been subjected to considerable attack. Bennett Harrison argues that there is no difference in the quality of jobs held by blacks in and out of suburbs, that racial discrimination in the labor market may restrict opportunities to blacks even if they moved nearer to the jobs, and that dispersal of the ghetto may disperse the ghetto problem but would not improve black employment. Paul Offner and Daniel Saks show that Kain's support for his hypotheses is subject to considerable statistical bias. By a slight alteration in the functional form of the estimated equation, Offner and Saks are able to demonstrate that removal of housing segregation may result in a loss of black jobs rather than a gain as Kain suggests. Kain's response to this is that even if creating ghetto pockets in the suburbs is needed to move blacks closer to the jobs, then this policy is to be preferred over maintaining one "gilded ghetto." Joseph Mooney contends that spatial separation of jobs and residences plays a less important role in reducing employment opportunities than Kain asserts. Both Charlotte Freemon and Wilfred Lewis have questioned Kain's hypothesis and have examined the components of industrial relocation to show that employment opportunities have not declined for ghetto residents.

III. A Theoretical Perspective

The notion that spatial separation of jobs and residences diminishes employment opportunities suggests a job search theory of wage differentials. But such a theory could lead to just the opposite policy conclusion than that offered by defenders of ghetto dispersion. In the conventional job search model (for example, see Steven Lippman and John McCall), job searchers are assumed to be risk neutral, to maximize their expected net benefits, and decide whether or not to accept any given job

offer according to a reservation wage decision rule. If the wage offer equals or exceeds the reservation wage, accept the offer; otherwise, continue to search. One important property of the reservation wage is that it declines for higher search costs. Workers with higher reservation wages can be expected to search longer and thus be unemployed frictionally longer. The longer duration of search is expected to yield the benefit, however, of higher reemployment wages. Hence, policies like ghetto dispersal which might reduce search costs (by reducing travel costs of search and improving information flows via decentralization of residences) would lead to higher short-run unemployment among the beneficiaries of the policy. Longer-run gains from higher reemployment wages may not be forthcoming if cyclical fluctuations in economic activity dictate that employers use last-hired first-fired rules.

Another contradictory result flowing from a job search view of ghetto dispersal is that reservation wages may rise not only due to lower search costs, but because of higher expectations about the opportunities in suburban labor markets. If there are serious overestimates by workers of their potential earning prospects and if through time the reservation wage is revised downward only slowly, then one consequence of ghetto dispersal could be decentralized pockets of high unemployment throughout the metropolitan area among former ghetto residents.

Wage inequality resulting from spatial separation of jobs and residences suggests another view casting doubt on the ghetto dispersal argument. Suppose that search costs are negligible, perhaps because of the existence of a costless job referral service. But suppose that travel cost, increasing in the distance between job and residence, reduces utility of jobs accepted. The risk-neutral worker's utility could be given by wage income less travel costs. If the hours worked were identical for both ghetto residents who work in the suburbs and those who work in the inner city, and if the opportunity cost of time were identical for everyone, then the requirement that all workers' utilities be equal implies the well-known result that black workers living in the ghetto, who work in the ghetto, receive lower wages than black ghetto residents working in the suburbs. However, it is easy to see in such an abstract world how moving blacks to the suburbs need not make black workers better off, unless there really are better jobs there.

Of course, the theoretical arguments in favor of ghetto dispersal rarely adopt such simple reference models. The point of looking at a model of rational job choices is, however, to present a valid challenge. If ghetto dispersal cannot be expected to work in a simple abstract model, how could it be expected to work in a complex realistic model?

IV. Evidence from an Isolated Case

If there is racial discrimination in suburban labor markets, or if there are substantial employment losses as a result of decentralization of black residential communities, the ghetto dispersal argument appears to degenerate into merely dispersal of the ghetto problem.

In our earlier paper, limited evidence was presented suggesting that although there may not be discrimination in suburban labor markets, living and working in neighborhoods of similar levels of segregation appears to improve one's probability of employment. The evidence is from a study of low-income job applicants who were clients of a Comprehensive Employment and Training Act (*CETA*) job referral service in Baltimore during 1974–75. Higher mean wages were offered to both black and white clients in suburban jobs compared to wage offers in city jobs. Similarly, job offer probabilities were higher for suburban jobs than city jobs, suggesting the often-argued point that blue collar vacancies have grown faster in the suburbs than in the city. Although blacks are less likely than whites to obtain a job offer in the suburbs, the wages of placed blacks and whites are not significantly different. Important wage disparities exist between blacks and whites in the suburban construction industry, but the low proportion of total referrals accounted for by construction leave the overall mean wages for blacks and whites about the same.

In a limited dependent variable model predicting the probability of a job offer, controlling for age, education, race, duration of job search, and marital status, it is found that applicants who apply for jobs in the suburbs are more likely to obtain an offer than those who apply in the city. Moreover, those who travel away from the central business district to work can expect sixteen cents more per hour in higher wages on the average controlling for age, sex, race, education, duration of search, and distance travelled.

However, a different story is told by data on racial composition of census tracts. Controlling for distance, race, and a vector of other variables, *CETA* clients are less likely to show up for a scheduled interview in the suburbs (where census tracts are largely nonblack). Of those that show, applicants who interview in census tracts of the same level of racial segregation as their own neighborhood generally have higher probabilities of being offered a job. For example, applicants from home tracts 0–20 percent black are more likely to be hired in tracts 0–20 percent black than those from any other area. Similarly, those from nearly all black tracts have the highest odds of being placed in a job located in an all black tract. This evidence is not inconsistent with higher probabilities of job offers in suburban locations reported earlier. It corroborates the contention that even suburban residences are highly segregated. High wages, on the other hand, are concentrated in essentially white census tracts, regardless of the census tract of the residence. While there is a gain in wages offered in suburban areas, there are losses because blacks, wherever they live, are less likely to show and less likely to be hired if they do show.

V. Consequences of Urban Revitalization

Whether because of declining tax bases, deteriorating public services, the energy crunch making housing closer to the urban core more attractive to middle-income workers, or because of a general aesthetic lure back to the cities, there has been a renewed desire to make large urban areas comfortable, safe, and convenient places to live. Either by design or consequence, black inner-city residents have been displaced and new pocket ghettos are becoming the future problems of the suburban towns. For whatever reasons, the ghetto is being dispersed and it is difficult to see how poor blacks will benefit. However, more comprehensive evidence, emerging from the 1980 Census, may show that the scattered black residences, the isolated black pockets of what was once a community, indeed, have fared well.

The difficulty in assessing whether development of a centralized minority community in a large urban area is to be preferred to dispersing the residences closer to changing employment prospects goes beyond merely arguing over whether the dispersed prospects exist or are viable employment alternatives. The difficulty becomes one of comparing the welfare gains and losses of both the black community which loses a few positive externalities through dispersal, and the rest of the urban area which could obtain a few negative externalities from continued gilding of the ghetto. If factors such as crime really are more deleterious in monolithic concentrations of blacks near the urban core rather than dispersed throughout the metropolis, one must question whether the positive aspects of community given up, such as the political power generated from geographical proximity to a homogeneous constituency, outweigh the gains achieved in reducing the negative externalities.

REFERENCES

Charlotte Freemon, *The Occupational Patterns in Urban Employment Change, 1965–1967*, Washington 1970.

Bennett Harrison, *Urban Economic Development*, Washington 1974.

David T. Herbert, *Urban Geography—A Social Perspective*, New York 1973.

J. F. Kain, "Housing Segregation, Negro Employment and Metropolitan Decentralization," *Quart. J. Econ.*, May 1968, *82*, 175–98.

W. Lewis, Jr., "Urban Growth and Suburbanization of Employment: Some New Data,"

unpublished manuscript, Brookings Instit. 1969.

S. A. Lippman and J. J. McCall, "The Economics of Job Search: A Survey," *Econ. Inquiry*, Sept. 1976, *14*, 347–67.

J. D. Mooney, "Housing Segregation, Negro Employment and Metropolitan Decentralization," *Quart. J. Econ.*, May 1969, *83*, 299–312.

P. Offner and D. A. Saks, "A Note on John Kain's 'Housing Segregation, Negro Employment and Metropolitan Decentralization,'" *Quart. J. Econ.*, Feb. 1971, *85*, 147–61.

K. E. Phillips and S. L. Myers, "Job Search, Spatial Separation of Jobs and Residences, and Discrimination in Suburban Labor Markets," P-6189, Rand Corp. 1978.

Noncooperative Equilibrium and Market Signalling

By JOHN G. RILEY*

Among many recent developments in the economics of information, none has generated more controversy than the concept of "market signalling" introduced by A. Michael Spence. If buyers are less well-informed about product quality than sellers, and no additional information is available, market-clearing prices must reflect some weighted average of product quality. Then if potential sellers of the highest quality products have the greatest opportunity costs and these costs exceed price, they will not enter the market.

This "adverse selection" phenomenon is, however, offset if sellers of higher quality products can adopt activities that operate as a "signal" to potential buyers. Intuitively, an activity is a potential signal if entering into it has a lower marginal cost for sellers of higher quality products. For example, the fly-by-night operator faces a higher advertising cost per unit of sales than the new entrant who plans to build and then maintain his reputation. Or, in the labor market, productivity on the job is positively correlated with performance in school. Therefore, the higher productivity worker has on average a lower personal cost of obtaining a given set of educational credentials. Similarly, in purchasing insurance the marginal cost of accepting a higher coinsurance rate for a specific loss is lower for those with a lower probability of loss.

But on closer inspection, it turns out that "informational equilibria"—that is, equilibria in which signalling is needed to distinguish product quality—do not have the stability properties of classical Walrasian equilibria. In papers by Michael Rothschild and Joseph Stiglitz, and the author (1975, 1977) it has been shown that unless the difference between quality levels is sufficiently large there is, for every set of signal-price pairs S, some alternative $s' \notin S$ which, if offered by a single seller, would improve his lot. That is, there is no Cournot-Nash signalling equilibrium.

One possible inference that might be drawn from this result is that signalling could not be an important phenomenon in a competitive economy. After all, if any attempt at signalling were to result in interference, in the form of competitive responses, there would be no opportunity for individuals to identify the correlation between the level of the signal and the underlying quality of the product.

However this is too simplistic since the interference cannot take place until *after* signalling has been established. It is therefore more appropriate to suppose that an economy is initially in a state of informational equilibrium with signals reflecting product quality, and to ask whether the potential instability is likely to lead to a general collapse of the equilibrium.

Following the approach adopted by Charles Wilson there have been several related attempts to introduce quasi-dynamic considerations in which every agent takes into account the *reaction* by other agents when contemplating a "defection" from the initial set of signal-price pairs S. It is then argued that the set S forms a stable equilibrium if, for every alternative s' that would make a single defector better off, this same signal-price pair would make the defector worse off after the reaction by other agents.

Where these attempts differ is in the

*University of California-Los Angeles. This research was supported by National Science Foundation Grant SOC-76-13443 to the Rand Corporation.

assumptions made about the type of reactions that potential defectors anticipate. In this paper these differences are highlighted by considering a simple model in which there are two classes of agents who are not directly distinguishable. For concreteness the discussion is placed in the insurance context. A simple relabelling converts the model into one of educational signalling.

I. A Simple Insurance Model

Consider an economy in which all agents face some risk of incurring a loss of L dollars. Agents are identical[1] except in the probability of loss π, which may take on the values π_h or π_l ($\pi_h > \pi_l$). With costless information about π and a competitive insurance industry offering fair policies, all consumers would obtain full coverage. However, with π unobservable, insurance companies offer policies which include a coinsurance rate y. Then if a loss occurs a company pays out only $(1 - y)L$ dollars.

Associated with any coinsurance rate y^* is a payout-premium ratio R^*. Suppose only consumers in the ith risk class purchase the policy $s^* = \langle y^*, R^*\rangle$. The expected payout divided by the premium is then $\pi_i R^*$. Neglecting costs and assuming that entry into the insurance industry continues until expected profit is zero, that is expected payout equals premium revenue, we require $\pi_i R^* = 1$.

More generally, if $\bar{\pi}$ is the average probability of loss for those purchasing s^*, the zero expected profit condition becomes $\bar{\pi} R^* = 1$ or $R^* = 1/\bar{\pi}$. In general companies will offer a different payout-premium ratio R for a different coinsurance rate y. Then each expected utility maximizing consumer selects a level of y yielding the solution of

$$\max_y U(\pi; y, R(y)) = (1 - \pi)u(I - p(y)) + \pi u(I - yL - p(y))$$

where $p(y) = (1 - y)L/R(y)$, I = income and u = a standard utility function. Note that yL is the level of coinsurance so $(1 - y)L$ is the insurance coverage and $p(y) = (1 - y)L/R(y)$ is the insurance premium (assumed paid whether or not a loss occurs). Indifference curves in $\langle y, R\rangle$ space are depicted in Figure 1 for the two risk classes. There are two important aspects of these curves. First, for every policy $s = \langle y, R\rangle$ the slope of the indifference curve for the high risk class ($\pi = \pi_h$) is greater. That is, the increase in the payout-premium ratio necessary to maintain expected utility in the face of a higher coinsurance rate is always greater. Second, suppose a consumer is offered a fixed payout-premium ratio R^* but is free to select his coinsurance rate. Then if the payout-premium ratio is unfair, i.e., $R^* < 1/\pi$, the consumer will prefer some coinsurance. The intuition is straightforward: With fair insurance a consumer will shed all risk. However a decline in the payout-premium ratio raises the cost of each unit of coverage so less than full coverage is preferred. At some point the payout-premium ratio drops so low that no insurance is purchased ($y = 1$).

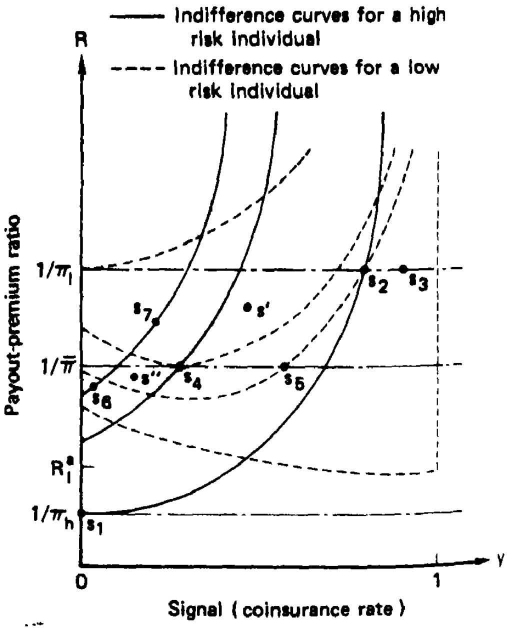


FIGURE 1. PREFERENCES OF THE TWO RISK CLASSES

[1]With observable differences the only change in the conclusions is that the results are contingent upon an agent being in some known class. Adding unobservable differences generates noise, but again does not change the conclusions.

Consider for example the indifference curves of the low-risk class in Figure 1. For $R > 1/\pi_l$ the indifference curves are upward sloping everywhere so, holding R constant, the most preferred coinsurance rate is zero. For $R < 1/\pi_l$ the indifference curves are first downward sloping so, holding R constant, some coinsurance is preferred. Finally for $R \leq R_l^o$ the payout-premium ratio is so low that the low-risk class prefer to purchase no insurance ($y = 1$).

II. Alternative Equilibrium Concepts

Spence in his initial analysis (1973, 1974) describes a set of contracts S as an equilibrium set if, when agents select freely among these contracts, each such contract breaks even. Here we shall call such a set *informationally consistent* if the different risk classes accept different contracts and *weakly informationally consistent* if one or more contract attracts both classes.[2] It is easy to see that there is a whole family of informationally consistent sets of contracts. For example in Figure 1 $\{s_1, s_2\}$ and $\{s_1, s_3\}$ are both informationally consistent. In each case the high-risk class purchases full coverage and the payout-premium ratio ($1/\pi_h$) generates zero expected profits. The low-risk class signals by purchasing a policy with a positive coinsurance rate. Another informationally consistent set of policies is the one element set $\{s_1\}$. Once again the high-risk class purchases full coverage but now the low-risk class is better off without any insurance. This is an illustration of George Akerlof's adverse selection equilibrium with only the "lemons" left in the insurance market.

There is also a family of weakly informationally consistent policies of which s_4 and s_5 are members. In each case the single policy attracts both risk classes and the zero expected profit condition is satisfied ($R = 1/\bar{\pi}$).

It therefore appears as though multiple equilibria are the rule rather than the exception in a world of informational asymmetry. However when the stability of these equilibria are examined the problem is not whether there are too many, but whether there are any. To be precise, unless differences between risk classes are sufficiently large, none of the "informational equilibria" satisfy the requirements of a Cournot-Nash equilibrium.

To see this consider first the weakly informationally consistent policy s_4. If a new firm enters and offers the alternative policy $s' = \langle y', R' \rangle$ it attracts only the low-risk class. Since the payout premium ratio R' is less than $1/\pi_l$ the new policy is profitable. Next consider a set of contracts that are informationally consistent, for example, $S = \{s_1, s_2\}$. If a firm defects from this set offering the alternative policy $s'' = \langle y'', R'' \rangle$ it attracts both risk classes and yields expected profits since R'' is less than $1/\bar{\pi}$.

However note that while the defector offers s'' it is always possible for a second insurance company to react with an offer such as s'. The defector, instead of making expected profits finds that the low-risk types have been skimmed off by the reactor. The latter now makes profits while the defector loses money. Note furthermore that since the reactor makes profits on all those accepting his policy the worst that could result from additional policy offerings by other firms is that these profits are eliminated. The reactor therefore stands only to gain from his new offering.

But if each agent recognizes that these opportunities for profitable reaction exist it seems reasonable that they will effectively deter a contemplated defection. Elsewhere I have shown, in a more general context, that under the Spencian assumptions there is always a unique *reactive equilibrium* in which every potential additional policy is open to this threat of reaction. For the two-class case depicted in Figure 1 this is the pair $S_R = \{s_1, s_2\}$. Note that any pair of informationally consistent policies must lie respectively on the horizontal lines $R = 1/\pi_l$, $R = 1/\pi_h$. Policy s_1 is therefore best for the high risk group.

[2]Whenever the high risk class is indifferent between two insurance policies we assume that it selects the lower coinsurance rate. This is a natural assumption since any significant amount of switching would lower the break-even payout-premium ratio on the policy with higher coinsurance and so make it strictly less desirable to the high risk class.

Given s_1, the best policy for the low-risk class that distinguishes the two classes is s_2. Thus the set S_R is a Pareto optimal informationally consistent set of policies.

However, returning to Figure 1, it can be seen that the single policy $s_4 = \langle y_4, R_4 \rangle$ is strictly preferred to the reactive equilibrium set of policies by both risk classes. Moreover s_4 lies on the horizontal line $R = 1/\bar{\pi}$ so it is a weakly informationally consistent policy. Since any other policy on this horizontal line must yield a lower expected utility to at least one risk class s_4 is Pareto optimal among the set of weakly informationally consistent policies.

Wilson has shown that if firms respond to a defection from some set of policies S by dropping just enough policies so that those remaining at least break even, the resulting equilibrium is unique and is Pareto optimal among weakly informationally consistent sets of policies. In the simple two-class model examined here this equilibrium is the single policy s_4.

As we have already seen, if insurance companies begin offering only s_4 it is profitable for one company to drop s_4 and offer instead a policy like s'. This skims off the profitable lower-risk class and generates expected losses for all the other insurance companies. But if the latter all respond by dropping their initial policy, s' will be chosen by both risk classes. Since $R' > 1/\bar{\pi}$ the defector now has expected losses.[3]

Wilson also considered, but did not pursue, the possibility that insurance companies might subsidize one kind of policy with the profits from another policy. This idea has since been followed up by Hajime Miyazaki for the two-class case and developed more fully by Spence (1977).

For our illustrative model it can be shown that there are policy pairs such as $\{s_6, s_7\}$ in Figure 1 with the low-risk class subsidizing the high-risk class and the two policies breaking even in the aggregate. It can be seen that both risk classes would prefer the pair $\{s_6, s_7\}$ over s_4. Spence has shown that for a general n-class model, where firms react by dropping loss-making policies, there is a unique equilibrium which is weakly informationally consistent in the aggregate. He has also established that this equilibrium is Pareto optimal among all sets of contracts satisfying aggregate consistency.

In Figure 1 the difference between the two risk classes is such that all three equilibria are distinct. This is not necessarily the case Indeed if differences among risk classes are sufficiently large the three equilibria will coincide. However the converse is also true. Whenever these differences are sufficiently small the equilibrium achieved is necessarily sensitive to the assumption made about how firms will respond to a defection. Since the three equilibria are Pareto optimal over successively larger sets of insurance policies, the equilibria are themselves Pareto ranked It is therefore of considerable importance to focus more closely on the differences in assumptions, in order to better understand how well the marketplace allocates resources when there is informational asymmetry. My own admittedly interested view is that the more favorable the equilibrium in the Pareto sense the harder it is to justify it as being "competitive."

For example, in the Miyasaki-Spence equilibrium insurance companies offer a menu of policies in which the higher risk classes are subsidized by the lower risk classes. It is therefore possible for an insurance company to make money simply by turning away some applicants for the policies accepted by the higher risk classes. Apart from the difficulty of detecting this form of rationing I find it hard to swallow the assumption that *all* other firms would respond with a threat to drop policies.

In the equilibrium proposed by Wilson each insurance policy generates zero expected profits so any "cheating" must take the form of an announcement of a new type of policy The detection issue is therefore not so serious.

[3]Herschel Grossman has pointed out that, in principle, Wilson's equilibrium is achievable via the reactions of *buyers* rather than *sellers* of insurance. When the new policy s' is offered, the high-risk class recognizes that the departure of the low-risk class will force insurance companies to lower the initial payout-premium ratio to $1/\pi_4$. The new policy s' is therefore preferred by both risk classes.

However there remains a distinct flavor of collusiveness about the envisaged response. For example in the two-class model all insurance companies respond to any defection by dropping out of the insurance business. More generally, *all* companies drop policies until those remaining at least break even.

In contrast the "reactive equilibrium" relies on a threat not from the market participants as a whole but from only one other firm (reactions by more than one firm only serve to strengthen the equilibrium). The key assumption is that every firm believes that at least one competitor is maintaining a close watch on its "product line" and will exploit any sure gain.

III. Concluding Remarks

In the Walrasian or Arrow-Debreu equilibrium under uncertainty traders sign contracts which are contingent upon the eventual state of the world. The central point of this paper is that serious difficulties arise when an attempt is made to extend the Walrasian equilibrium to a world of uncertainty without markets for every contingency.

It is important to recognize that the concept of an informationally consistent set of contracts, as first suggested by Spence, is a natural extension of the Walrasian approach. Only the consumer initially knows the probability of loss and hence the value of the risk that he is trading to an insurance company. Insurance policies are then characterized (labeled) by the level of the signal (the coinsurance rate) y. Each consumer, facing the same set of parametric payouts per dollar of premium $R(y)$, chooses to trade that risk which maximizes his own expected utility. On the other side of the market insurance companies, acting as price takers, have beliefs $\pi^*(y)$ as to the probability of loss associated with each level of y. Then if the set of policies is informationally consistent these beliefs are correct and expected profits are zero ($\pi^*(y)R(y) = 1$ for all y).

The recent research summarized above has demonstrated that informational consistency is not sufficient to eliminate opportunities for potential gain. Indeed in general there is no Cournot-Nash equilibrium. However it has been argued that the implied potential instability is not, after all, devastating. Instead, by building into the equilibrium concept a recognition of possible reactions by other agents, stability is achieved. The transfer of information via markets can therefore be explained as a noncooperative equilibrium phenomenon. It will be of considerable interest to see whether the new equilibrium concepts also prove useful in other cases where the Cournot-Nash approach fails.

REFERENCES

G. Akerlof, "The Model for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quart J. Econ.*, Aug. 1970, *89*, 488–500.

H. I. Grossman, "Adverse Selection, Dissembling and Competitive Equilibrium," work. paper, Brown Univ., Dec. 1977.

H. Miyazaki, "The Rat Race and Internal Labor Markets," *Bell J.*, Autumn 1977, *8*, 394–418.

J. G. Riley, "Competitive Signalling," *J. Econ. Theory*, Apr. 1975, *10*, 175–86.

———, "Informational Equilibrium" Rand Corp., no.R-2059-NSF, Apr. 1977.

M. Rothschild and J. E. Stiglitz, "Equilibrium in Competitive Insurance Markets: The Economics of Imperfect Information," *Quart J. Econ.*, Nov. 1976, *90*, 629–49.

A. Michael Spence, *Market Signaling: Information Transfer in Hiring and Related Processes*, Cambridge, Mass. 1973.

———, "Competitive and Optimal Responses to Signals," *J. Econ. Theory*, Mar. 1974, *7*, 296–332.

———, "Product Differentiation and Consumer Choice In Insurance Markets," disc. paper no. 585, Harvard Inst. Econ. Res., Nov. 1977.

C. A. Wilson, "A Model of Insurance Markets with Incomplete Information," *J. Econ. Theory*, Dec. 1977, *16*, 167–207.

Equilibrium and Agency—Inadmissible Agents in the Public Agency Problem

By STEPHEN A. ROSS*

An agency relationship arises when one party, the agent, takes actions on behalf of another party, a principal. The theory of agency was first developed in different contexts by Robert Wilson (1968, 1969), A. Michael Spence and Richard Zeckhauser, and the author (1973, 1974), and has been extended in a number of different directions. Steven Shavell, for example, has examined moral hazard issues associated with imperfectly monitoring the agent's effort and Milton Harris and Artur Raviv have looked at optimal contract structures. This paper takes up some different issues motivated by the desire to study the qualitative properties of equilibrium in markets for agents.

I

Consider a market in which agents compete to serve the interests of principals. Agents have a sure opportunity cost or wage ω_o, and an agent will supply his services to the market as long as the certainty equivalent wage of serving as an agent does not fall short of ω_o. The service performed by agents consists of choosing an act $\alpha \in A$, a set of feasible actions. Subsequent to the choice, a state of nature $\theta \in \Omega$ is realized yielding a monetary payoff, $w(\alpha, \theta)$. The payoff is shared, with the agent receiving a fee schedule $f(w(\alpha, \theta))$, and the principal receiving $w - f$. By assumption the fee schedule is a function only of the payoff and is not directly state dependent. (This constraint most commonly arises from the assumption that θ is unobservable or if observed is not verifiable by some parties to the contracts.)

The contract or fee schedule, $f(\cdot)$, is the object of equilibrium in the market and an equilibrium occurs when a set of fee schedules has been found with the property that the demand for agency services by principals equals the supply. Assume that agents act to maximize the expected utility of the fee they receive, $\max_{\alpha \in A} E\{A(f(w(\alpha, \theta)))\}$, and that principals seek to maximize over both their choice of an agent and their fee schedules $\max E\{C(w - f(w))\}$, where A and C are the von Neumann-Morgenstern utility functions of both the agent and principal, respectively. For the moment, principals and agents are assumed to share common information, the same probability measure on Ω, and, in addition, principals monitor agents to the extent of knowing what it is that they truly maximize.

In this market principals search over agents for the best (fee, action) or equivalently (fee, strategy) pair, and agents compete over these pairs. With agents in elastic supply, competition will insure that no rents can be earned, that is, for agents active in the market, the return will be bid down to the opportunity wage. In equilibrium a single agent type may serve many principal types. (Conversely, of course, many agents may serve a single principal, but in our model this case is somewhat singular.)

The striking feature of the simple market I have set up is that a less risk-averse (less conservative) agent will always be able to outbid more risk-averse agents in servicing *any* risk-averse principals! The following theorem formalizes the absolute advantage of less conservative agents.

THEOREM 1: *Suppose that agent A is more risk averse than agent B, and that the supply of such agents is elastic at the same opportunity wage, ω_o. In equilibrium, no type A agents will be in the market.*

*Professor of organization, management, and economics, School of Organization and Management and department of economics, Yale University. I am grateful to NSF Grant #SOC77-22301 for support.

PROOF:

Adopting the traditional Arrow-Pratt measures of risk aversion, since A is more risk averse than B, there exists a monotone, concave transform $G(\cdot)$, such that $A(\cdot) = G(B(\cdot))$ (see John Pratt). If A types are supplying the pair (f_a, α_a) we have $G(E\{B(f_a(w(\alpha_a, \theta)))\}) > E\{G(B(f_a(w(\alpha_a, \theta))))\} = E\{A(f_a(w(\alpha_a, \theta)))\} \geq A(\omega_o)$, which implies that $E_\theta\{B(f_a(w(\alpha_a, \theta)))\} > B(\omega_o)$.

Thus, type B agents would be willing to supply the same service, that is, the same (fee, act) pair, and since the inequality is strict, by lowering the fee by an arbitrarily small amount (in states where it is positive) type B agents will outbid type A agents. In equilibrium, since type B services are elastically supplied no type A agents will survive in the market.

Theorem 1 verifies that in a world where monitoring is complete and contracts are perfectly enforced more conservative agents will not be able to compete. Less risk-averse agents are willing to take greater chances and will offer the principal more risk offset. There are, of course, several assumptions behind such a result, and some of them can be relaxed with obvious effects. To begin with, if the supply of agents is not perfectly elastic, then in equilibrium more conservative agents may still be in the market and less risk-averse ones will earn Ricardian rents analogous to the rents earned on a specific factor by a firm with a resulting low-cost curve. The most conservative agent in the market will be at the margin and just earn the opportunity wage. In addition, in general agents may not be nicely ordered in terms of risk aversion and each agent type may have a particular market niche.

Theorem 1 does not depend on the assumption that principals and agents have identical information sets. The next theorem explicitly makes this point.

THEOREM 2. *Suppose that the conditions of Theorem 1 hold with the modification that the principal does not possess the same information set as the agents. Rather, the principal assigns the same prior distributions over possible information sets to both agents. It follows, once again, that no type A agents will be in the market.*

PROOF:

The proof is identical, since at the same (fee, action) pair the principal will still be indifferent to which agent is employed and if agent B offers a dominating fee, agent A will not be employed.[1]

II

In general, then, conservative agents can survive in the market only because of costs in monitoring and enforcing agents' actions; moral hazard is a necessary condition for conservative agents to be used. With moral hazard present, a principal faces a tradeoff in choosing an agent. The less risk averse the agent, the more willing he is to share in the total risk of the payoff, but as the agent becomes more and more risk neutral he will make increasingly risky decisions. This would imply, though, that a principal would never choose an agent who was more risk averse than agents who, in turn, were more risk averse than himself. To do so would result in both too conservative decisions and less risk sharing than could be obtained by using a less

[1]Notice that it is crucial to this result that both types of agents actually have the *same* information sets. If the more conservative agent had a finer partition than the less risk-averse one, then he might be able to successfully compete by offering a superior choice of action. Of course, if the type B agent knows that A possesses better information and can monitor A's superior choice, then by simply offering the same (fee,act) pair he can still outbid A, secure in the knowledge that if A finds such a pair acceptable he must also. Considerations of this sort are beyond the scope of this paper. Artur Raviv has pointed out that if an agent is risk neutral, then he can offer the principal a constant claim. Such a perfectly insured contract would dominate any other offer and is not dependent on the information set. Whether such feasible, uniformly dominating, information free contracts exist in the case of Theorems 1 and 2 is a conjecture. Such contracts do, however, assume that agents possess resources external to the payoff, and that their commitments to hedge against adverse payoffs can be bonded. The above proofs were specifically constructed to not rely on such a condition and can, therefore, be applied in situations where agents cannot fully insure principals' risk—a characteristic of many agency problems.

risk-averse agent who was still more conservative than the principal. To examine these phenomena it is useful to consider a somewhat altered problem where the risk-sharing dimension is not present and where the focus is on the compatibility of the choice of an action.

Let us assume that agents act to maximize their expected utility at a given fee schedule and that principals receive the whole payoff $E\{C(w(\alpha, \theta))\}$, and maximize this over their choice of an agent. This could be thought of as a public choice problem where an agent—perhaps a political agent—chooses, and the principal receives a public good return. We will call this the public agency problem. In performing this maximization it will be assumed that principals do know the preference structure of agents.

Since, in this context, the payoff is not shared with the agent, in choosing an agent the principal is interested only in picking someone whose preferences are in some sense closest to his own. In the limit, the principal would prefer an agent whose preferences were similar under the fee schedule, that is, an agent for whom there were constants $a > 0$ and b with $A(f(\cdot)) = a\,P(\cdot) + b$; such an agent would choose in all situations exactly as would the principal (see the author 1973, 1974).

Even without similarity, though, it may still be that principals prefer the net ordering of one agent to that of another at a given fee. This is the problem of admissibility on which I will focus in the remainder of the paper; when is it the case that agents can be ordered by principals independent of the particular decision problem at hand. For the moment, we will ignore the fee schedule and will subsume it into the agents' utility functions. Formally, we are interested in characterizing those triples (A, B, C) with the property that the principal C always agrees with agent B in any disputes between A and B. In such a situation we will say that C prefers agent B to agent A and that A is inadmissible. Of course, if C and B always agreed in *all* situations then they would have identical preferences. We only require that when A and B disagree, C would rather have B's choice than A's. As might be expected, this situation arises only under very limited circumstances, and the following result shows that this must imply that A, B, and C are linearly related.

THEOREM 3: *Suppose that C agrees with B whenever A and B are in disagreement in a binary choice between two lotteries. There must then exist a semipositive vector p such that (up to an affine transform)* $p_aA - p_bB + p_cC = 0$.

PROOF:

To simplify the mathematics we will assume that the utility functions are defined on a discrete n-point equally spaced lattice. Now, each can be represented as a vector and lotteries will simply be probability vectors on this lattice. If x and y are any two probability vectors on the lattice then defining $\Pi \equiv x - y$, the formal statement of the theorem is that for all Π whose components sum to zero, $\Pi' a > 0$ and $\Pi' b > 0$ imply that $\Pi' c \leq 0$, where a, b, and c are the respective vector representations of the utility functions A, B, and C.

By an application of Farkas' lemma this will be the case if and only if there exist constants (p_a, p_b, p_c, p_e) with $p_a, p_b, p_c \geq 0$ and $p_a a - p_b b + p_c c + p_e e = 0$, where $e' \equiv \langle 1, \ldots, 1 \rangle$.

COROLLARY 1: *Suppose that C prefers B to A and that there is some pair of lotteries on which A and B disagree, then there exists a constant* $\lambda \geq 0$ *such that (up to an affine transform)* $C = B - \lambda A$. (The proof is straightforward.)

The next theorem pursues the question of whether by restricting the choice space of lotteries we might obtain less constrained results on preferences. Let us further restrict our attention to monotone, concave utility functions (keeping in mind that these are *cum* fee). Let L denote the unit lower triangular matrix,

$$L \equiv \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 1 & & 1 \end{bmatrix}$$

If a denotes a monotone increasing concave vector, then by letting α denote a second

difference term, with $\alpha_i \geq 0$ for $i = 2, \ldots, n-1$, we can build up a by using L and $a = LL'\alpha$.

The restrictions we wish to place on lottery choice is that in some intuitive sense lotteries differ only in return and risk. In formal terms, for any two lotteries x and y in the choice set one, y, say, was obtained from the other x, by shifting the x distribution in some direction and by adding a noise term. In distributional terms, y is distributed as $x + z + \epsilon$, where z is a nonnegative or a nonpositive random variable and ϵ is a noise term, with mean zero conditional on $x + z$. Clearly, if z is negative, then y is superior to x for all monotone concave agents since y is obtained from x by shifting some mass downward and by adding a mean preserving spread, ϵ. A more interesting case is where z is positive since now we have a tradeoff between extra return and risk.

We can represent such lotteries by again using the matrix L. The vector z will be a vector such that $x + z$ is still a probability vector and is preferred to x by all monotone utility functions. We can represent all such functions as Lu, with $u \geq 0$, and hence we must have $z'Lu \geq 0$ for all $u \geq 0$. This implies that z satisfies $z'L \geq 0$ and $e'z = 0$, the latter equality following from the requirement that $x + z$ be a probability vector.

Similarly, the noise term has the property that for all concave utility functions it lowers utility, i.e., for all α with $\alpha_i \geq 0$, $i = 2, \ldots, n-1$, $\epsilon'LL'\alpha \leq 0$, or $\epsilon'LL' \leq 0$ with $\epsilon'e = 0$. (It follows that the last component of $\epsilon'LL'$ must also be zero.) We can now prove the following theorem.

THEOREM 4: *Let S denote a set of pairs of lotteries generated as above, and let (A, B, C) be a triple of monotone, concave utility functions. The principal C will prefer agent B to agent A if and only if there exists two semipositive vectors p and q and two monotone decreasing concave functions G and H such that $p_aA - p_bB + p_cC = G$ and $q_aA - q_bB + q_cC = -H$.*

PROOF:

Again, to simplify the mathematics, let us consider a discrete problem. As before, the statement of the theorem implies that the equation systems $[z'LL' + \epsilon'LL'][\alpha, -\beta, \gamma] > 0$ or < 0, with $\alpha, \beta, \gamma \geq 0$, not have a solution for any (z, ϵ) satisfying the conditions in the text.

Since L is of full rank, $v' \equiv z'L \geq 0$, is an arbitrary semipositive vector whose first element is zero (by $z'e = 0$). Similarly, $s' \equiv -\epsilon'LL' \geq 0$, is also an arbitrary semipositive vector whose first and last elements are zero. Hence, for all such (v, s), $[v'_\alpha L' - s'][\alpha, -\beta, \gamma] > 0$ must have no solution. By a standard separation argument there exists a semipositive vector p, with $[v'L' - s'][\alpha, -\beta, \gamma]p \leq 0$ for all admissible (v, s) pairs. Since s can be arbitrarily large we must have $[\alpha, -\beta, \gamma] \geq 0$ for all components except the first and the last, and since v can also be large we must have $L'[\alpha, -\beta, \gamma]p \leq 0$, except in the first component. By similar reasoning there exists a vector $q \geq 0$ such that $[\alpha, -\beta, \gamma]q \leq 0$ and $L'[\alpha, -\beta, \gamma]q \geq 0$. Hence,

$$p_\alpha\alpha - p_\beta\beta + p_\gamma\gamma \equiv \delta; \quad \text{with } L'\delta \leq 0;\ \delta_2, \ldots, \delta_{n-1} \geq 0$$

$$q_\alpha\alpha - q_\beta\beta + q_\gamma\gamma \equiv \phi; \quad \text{with } L'\phi \geq 0;\ \phi_2, \ldots, \phi_{n-1} \leq 0$$

Multiplying through by LL' and defining G and H in an obvious fashion we obtain the result.

Interpreting Theorem 4 is tedious due to the different patterns of zeros that are possible. The following is one of the most important cases.

COROLLARY 2: *Agents A and B will always agree on S if and only if there exist concave and decreasing G and H with $p_aA = p_bB + G$ and $q_aA = q_bB - H$. It follows that $A = B$ (up to an affine transform).* (The proof is straightforward.)

The conditions of Corollary 2 suggest the following definition. We will say that A is more risk averse than B, ARB, if there exists $\lambda > 0$ and a decreasing concave function, G, such that $A = \lambda B + G$. (Notice that if ARB, then A is more risk averse by the usual Arrow-Pratt coefficient of risk-aversion criterion, but, the converse is not true.) It is easy to show that if ARB, then when faced with a

choice between x and $y = x + z + \epsilon$, A will choose x if B chooses x. The following specialization of Theorem 4 is the central result to which we have been building in this section; it confirms our earlier intuition.

THEOREM 5: *Given a principal C and two agents A and B, C will prefer B to A (over S) if either CRB and BRA, or ARB and BRC.* (In other words, if B is intermediate in risk aversion between A and C, then B will be the preferred agent.)

PROOF:

If ARB and BRC, then there exists $\lambda, \theta > 0$ and G, H concave and decreasing such that $A - \lambda B = G$ and $\theta C - B = -H$, which satisfies the conditions of Theorem 4. The argument is similar for CRB and BRA.

With this framework we could further explore what happens when agents of different types compete across fee schedules, but unfortunately the solutions to such equilibrium problems will generally be sensitive to the decision payoff structure even under the strong restrictions we have studied above. The following theorem, for example, dispenses with the intuition that when agents compete by offering identical fee schedules they can be simply ranked by principals as in, for example, Theorem 5.

THEOREM 6: *Even when the choice domain is restricted to S, there exists no triple (A, B, C) for which C prefers B to A for all possible fee schedules. In particular, this is true even if $C(\cdot) = B(\cdot) \neq A(\cdot)$, and even if f is restricted to monotone concave, or convex, functions.*

PROOF:

The proof is straightforward, but tedious. The basic approach is to observe that from Theorem 4 we require that $p_aA(f(\cdot)) - p_bB(f(\cdot)) + p_cC(\cdot) = G$, and $q_aA(f(\cdot)) - q_bB(f(\cdot)) + q_cC(\cdot) = -H$. By differentiating these conditions twice we can taxonomically rule out possibilities.

III

It is probably fair to conclude this paper with the observation that while we have learned some interesting tidbits in the micro theoretics of agency problems, the harvest for equilibrium analysis has been rather lean. The basic lesson is that with moral hazard and except in special cases, we will not be able to order agents in a simple fashion—analogous to ordering cost functions in an industry study. The decision structure will intermingle with preferences to jointly determine the properties of equilibrium in agency markets. Perhaps, what is here may prove useful for such a study, or, may be of some interest by itself.

REFERENCES

M. Harris and A. Raviv, "Optimal Incentive Contracts with Imperfect Information," work. paper no. 70-75-76, Grad. School Ind. Admin., Carnegie-Mellon Univ., Dec 1977.

J. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr 1964, *32*, 122–36.

S. A. Ross, "The Economic Theory of Agency: The Principal's Problem," *Amer. Econ Rev. Proc.*, May 1973, *63*, 134–39.

———, "On the Economic Theory of Agency and the Principal of Similarity," in Michael Balch, Daniel McFadden, and S. Y. Wu, eds., *Essays on Economic Behavior Under Uncertainty*, Amsterdam 1974, ch. 8.

S. Shavell, "On Moral Hazard and Insurance," disc. paper no. 557, Harv. Inst. Econ. Res., June 1977.

A. M. Spence and R. Zeckhauser, "Insurance, Information and Individual Action," *Amer Econ. Rev. Proc.*, May 1971, *61*, 380–87.

R. B. Wilson, "On the Theory of Syndicates," *Econometrica*, Jan. 1968, *36*, 119–32.

———, "The Structure of Incentives for Decentralization Under Uncertainty," *La Decision*, Du Centre National De La Recherche Scientifique, Paris 1969.

Equilibrium and Adverse Selection

By CHARLES A. WILSON*

A common characteristic of a large class of markets is that one side of the market is more informed than the other about the properties of one of the goods being traded. In some instances, this presents no serious problem. If the informed agents deal on a regular basis with the less-informed agents (for example, local grocers, barbers), there may be little incentive for the informed agents to take advantage of their superior information. In other cases, the problem may be avoided if it is profitable for specialists (or some government agency) to provide the information at a relatively low cost (for example, credit agencies, *Consumer Reports*). Frequently, however, these kinds of market responses provide at best a partial reduction in the informational asymmetry. There may still be substantial benefits to the less-informed agents from acquiring more information.

How the market will respond under these circumstances has been the focus of much recent research. Most of the attention, however, has been directed at examining the possibility that a signalling convention will emerge. The essential idea is that sellers of high quality products may choose contracts or invest in observable characteristics which distinguish their products from those of lower quality. Although I believe that signalling is an important and pervasive phenomenon, the conditions necessary for effective signalling to emerge may not always be satisfied. It is important, therefore, that we understand how the allocation of goods is affected in the absence of signalling, when the only variable that agents may use to distinguish quality is the price. This paper provides an overview of some of my recent research on this question.

My investigation begins with a welfare analysis of the Walrasian equilibrium. Specifically, the question is whether or not it is necessarily desirable for trade to take place at a price which clears the market. My analysis indicates that it is not. Under some conditions, it may be possible to make every agent in the market better off simply by raising the price. Besides generating some obvious policy implications, this result also suggests that the Walrasian equilibrium may not always be the appropriate equilibrium concept for this model. In a market with homogeneous goods, it is generally argued that independently of how the prices are set, as long as there is a large number of buyers and sellers, competitive pressures will force the price toward a stable Walrasian equilibrium. When an adverse selection problem appears, however, the possibility that some buyers may prefer a price higher than the one which clears the market casts some doubt as to whether such pressures will still be present. It is no longer obvious that the market will clear or even that all trade will take place at a single price.

These points can be conveniently illustrated using George Akerlof's model of the used car market. There is a set of cars of varying quality q distributed over an interval $[q_1, q_2]$ with density $f(q)$. Each agent in the economy has an identical utility function $u(c, q; t) = c + tq$ where c is consumption of other goods, q is the quality of car he consumes, and t is a parameter equal to his marginal rate of substitution of car quality for consumption. (If an agent does not consume a car, q may be set equal to zero.) The set of agents can be divided into two subsets, those that initially own exactly one car and those that own none. Each owner has the same utility parameter, $t = 1$; for the nonowners, however, t is distributed continuously over some interval $[t_1, t_2]$ with density $h(t)$.

As long as each owner can directly identify the quality of his own car, the supply curve will have the usual positive slope. A utility maximizing owner with a car of quality q will sell at price p if and only if $q \geqq p$. As the price rises, therefore, more cars will be supplied. If

*Department of economics, University of Wisconsin. This research was supported by the National Science Foundation under Grant SOC-77-08568.

we assume that nonowners can observe only the *average* quality of the cars sold at each price, however, the problem of adverse selection appears. The benefit from buying a car depends not only on the price, but also on the average quality of the car sold at that price. Given an average quality function $q^a(p)$, a utility maximizing nonowner with utility parameter t will then choose to purchase a car at price p if and only if $p < tq^a(p)$. It follows immediately, therefore, that the demand curve need not be downward sloping. If the supply elasticity of the average quality with respect to price is greater than one, the number of nonowners who choose to purchase a car actually increases with the price. Consequently, the price which equates supply and demand may not even be unique. An example is illustrated in Figure 1.

Price is measured on the vertical axis, the quantity of cars is measured on the right-hand side of the horizontal axis, the quality of cars on the left-hand side. The supply curve is labelled $S(p)$. The quantity supplied is zero for all prices less than q_1, then begins to rise as more cars are continuously supplied until at $p = q_2$, all cars are supplied and the supply curve becomes vertical. The particular shape of the supply curve depends on the density of cars $f(q)$.

To construct the demand function, it is useful to consider the preferences of the different buyers in the space of price and average quality. An indifference curve for an agent with utility index t is a straight line with slope t. Since by not purchasing a car any nonowner can attain the utility level associated with a zero car quality at a zero price, the set of buyers at any price is composed of those nonowners whose indifference curves through the corresponding point on the average quality function intersect the price axis at or below the origin. Consequently at any price p, the slope of the ray from the origin to the point $(q^a(p), p)$ represents the utility index of the marginal buyer. Those nonowners with a utility index higher than $p/q^a(p)$ will buy at price p; those with a utility index lower than $p/q^a(p)$ will not.

Because all sellers have the same utility index at any price, the quality of the marginal car always exceeds the average so that the average quality function must be monotonically increasing. In this example, however, I make an additional assumption. From price p to p'', the density of the higher quality marginal cars is sufficiently high that not only does the average quality of the cars increase, but also the ratio of average quality to the price. Consequently, the demand curve must be upward sloping over this interval. Assuming that the number of nonowners with a utility index between $p''/q^a(p'')$ and $p'/q^a(p')$ is also sufficiently large, a demand curve can then be constructed which intersects the supply curve at three different prices.

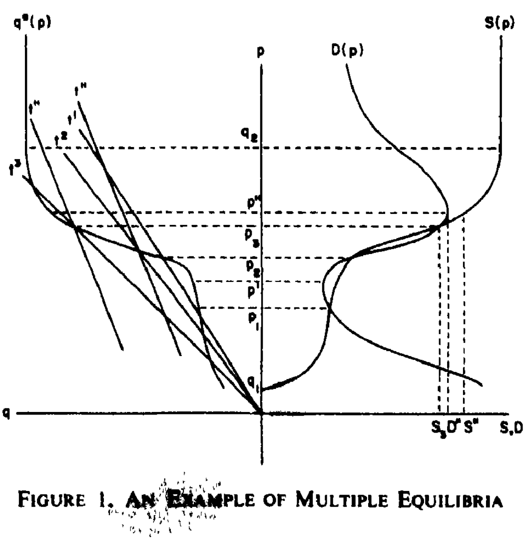


FIGURE 1. AN EXAMPLE OF MULTIPLE EQUILIBRIA

Now consider the welfare properties of these equilibria. Since the quality of their car does not depend on the price, owners always prefer the higher of any two prices at which they can find a buyer. Consequently, the higher is the equilibrium price, the higher is the welfare of each seller. A similar conclusion holds for the buyers, although the argument is less straightforward. To see this, consider the effect of increasing the price from p_2 to p_3. At p_2, the slope of any buyer's indifference curve is greater than or equal to t_2. When the price is increased to p_3, demand must have increased. This implies that the indifference curve of the marginal buyer has become flatter. Consequently, any buyer who demands a car at p_2 reaches a higher indiffer-

ence curve at price p_3. It follows, therefore, that p_3 is Pareto preferred to p_2.

If we explicitly introduce the possibility of rationing, this argument can sometimes be extended to generate a Pareto ranking across prices which are not necessarily market clearing. As long as there is no excess demand, the criteria for a welfare increase among all buyers remains unchanged. A price increase is preferred by all buyers if and only if the quantity demanded is higher at the higher price. An additional restriction, however, is now required to ensure that all sellers prefer a higher price.

When a price increase generates excess supply, sellers face a tradeoff between a reduction in the probability of selling a car and an increase in the price of the cars actually sold. In order for sellers to be better off, therefore, the probability of selling must not fall too rapidly as the price rises. In this model, all sellers are risk neutral; consequently as I have shown (1978), a sufficient condition for all sellers to prefer a higher price is that the percentage decrease in the probability of selling be less than the percentage increase in the price. This translates to the requirement that the difference between the elasticity of supply and the elasticity of demand be less than one. In Figure 1, demand is increasing from p_3 to p''. Therefore, p'' is Pareto preferred to p_3 if $S'' - D''/p'' - p_3 < S_3/p_3$.

As I noted in the introduction, the possibility that buyers may prefer a higher price suggests that competitive pressures may no longer force the market price to a Walrasian equilibrium. To illustrate this point, I will consider two extreme paradigms of a price-setting convention. In the first, the buyers are the price setters and the sellers price takers; in the second, the convention is reversed, sellers set the prices from which buyers choose to purchase. I have analyzed both these conventions in detail elsewhere (1977). Here, I will confine my attention to a brief sketch of the arguments and an outline of the results.

Suppose first that buyers are the price setters. In order to purchase a car, each nonowner must announce the price at which he will accept an offer to sell. Once a price is announced, it cannot be changed; however, no buyer is required to purchase more than one car. Assuming that sellers may costlessly search until they have found a willing buyer or decide to leave the market and assuming that the excess supply at any price is always rationed at random, the average quality function will remain unchanged. Under this convention, the equilibrium may or may not be identical to a Walrasian equilibrium, depending on the preferences of the buyers.

Suppose, for simplicity, that there is only one market-clearing price. If this price is announced by every buyer, there will be no excess demand and hence no incentive for any buyer to lower the price. Whether or not it is an equilibrium, therefore, depends on whether or not some buyer has an incentive to raise the price. When some buyer does announce a higher price, the equilibrium will no longer be characterized by a single price. A distribution of prices will emerge, generally extending above and below the market-clearing price. At all but the lowest price, the buyers will attract an excess supply of cars. Starting at the highest price, sales will be rationed and some of the cars successively offered at the lower prices until each owner has sold his car or decided to keep it.

Now consider the opposite price-setting convention. Suppose it is the sellers who must announce the prices from which the buyers must choose the prices at which they will purchase. If there is excess supply at any price, sales are rationed at random, and the ratio of demand to supply gives the probability that an owner will be able to sell his car at that price. In this case, the analysis becomes a bit more subtle as the role of expectations moves to the forefront. On the one hand, the number of buyers who demand a car at each price depends on the quality of cars they expect to be offered at each price. On the other hand, the quality of cars offered at each price depends on the number of buyers sellers expect to demand a car at each price. Consequently, whether or not the expectations of the agents on one side of the market are confirmed depends upon the expectations of the agents on the other side, and vice versa. Because of this interaction, the equilibrium

can take on a number of different forms. Unlike the case when buyers set the price, there is always a pattern of expectations which are consistent with a Walrasian equilibrium. In addition, however, there is always another pattern of expectations consistent with a continuous distribution of prices. Although the proof of this result is rather technical, it is possible to present part of the argument graphically and to illustrate some of the properties necesary for such an equilibrium to exist.

An equilibrium is illustrated in Figure 2. Price is measured on the vertical axis and probability of selling on the left-hand side of the horizontal axis. The curve labelled $\Pi^*\Pi^*$ represents the probability of making a sale at each price. The curves labelled v_1, v', v_2 represent indifference curves for the owners of cars q_1, q', and q_2, respectively, each drawn tangent to the $\Pi^*\Pi^*$ curve to illustrate optimal price for each seller to choose. The critical property illustrated in this figure is that the indifference curves of the owners of higher quality cars are always flatter than those of the owners of lower quality cars. Consequently, given the same probability function, owners of higher quality cars will announce the higher prices. To see why this is true, recall that $(p - q)$ is net benefit from selling a car with quality q. Therefore if $\Pi(p)$ is the probability of selling a car at price p, then $\Pi(p)(p - q)$ is the expected benefit to

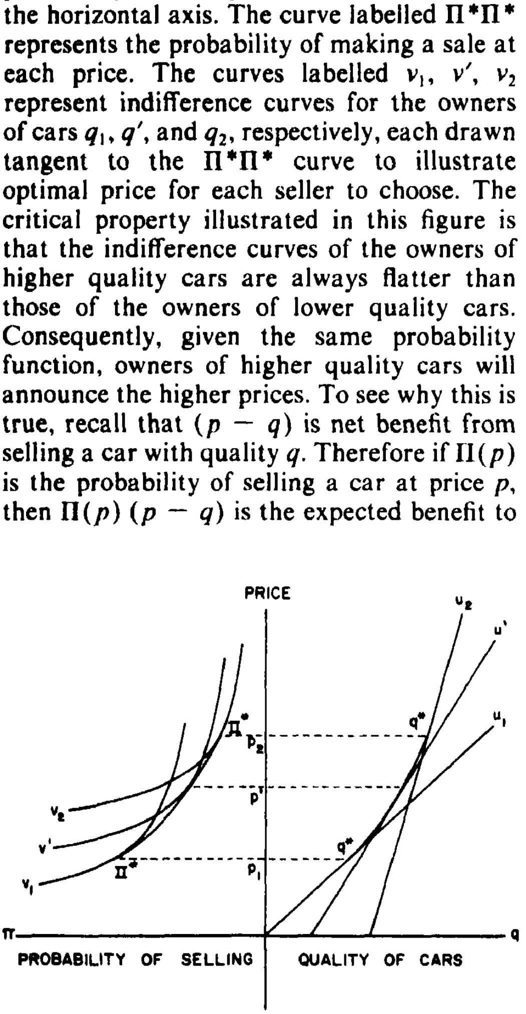


FIGURE 2. A DISTRIBUTION OF PRICES WHEN SELLERS SET THE PRICE

the seller from announcing price p. Since the percentage change in $(p - q)$ resulting from a unit increase in p is greater, the greater is q, it follows that owners of higher quality cars are willing to accept a lower probability of selling in order to obtain a higher price.

The resulting equilibrium quality function is illustrated on the right-hand side. Because the owners of higher quality cars announce higher prices it must be upward sloping. In addition, if buyers with different marginal rates of substitution of car quality for consumption are to choose distinct prices, it must be a convex function. What cannot be illustrated in this graph is that both of these functions can be adjusted to generate the equilibrium ratio of buyers to sellers at each price.

Let me conclude with a few remarks on the scope of these results. First of all, there is nothing essential in the analysis about the role of the uninformed agents as buyers. Since any buyer of a good must simultaneously be the seller of some other good, they might just as well have been called the sellers. What is important is that the conditions on the demand curve correspond to conditions on the demand of the *uninformed* agents. Similarly, the conditions on the supply curve must correspond to conditions on the supply of the *informed* agents. In financial markets, for instance, the firms typically play the role of the buyers and the consumers the sellers.

This leads to my second point. In markets where either the demand or the supply curve tends to be perfectly elastic, the results reported here are of little interest. For instance, in simple models of financial markets, it is typically assumed that the supply of contracts is perfectly elastic (not necessarily with respect to the size of the transaction with an individual consumer, but with respect to the number of consumers serviced). Although there may be multiple Walrasian equilibria in these models, it is never Pareto preferred to create an excess demand for contracts. Furthermore, the equilibrium will never be characterized by a distribution of prices.

Despite these limitations, I believe there is still a wide class of markets for which these results are of some interest, particularly

markets for the services of labor. I suspect that in many blue-collar occupations, there are examples of instances in which employers purposely offer a wage above the level which equates supply and demand in order to attract a higher average quality of workers. Some of the price dispersion in markets for consumer services may also be explained by the market response to the adverse selection problem. If, for instance, more conscientious doctors require more income per patient than do less conscientious doctors, it is possible that the better doctors may choose to charge more per patient and service fewer of them.

REFERENCES

G. Akerlof, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quart. J. Econ.*,, Aug. 1970, *89*, 488–500.

C. A. Wilson, "The Nature of Equilibrium in Markets with Adverse Selection," SSRI disc. paper no. 7715, Univ. Wisconsin, Nov. 1977.

———, "The Benefits of Price Maintenance in Markets with Adverse Selection," SSRI disc. paper no. 7817, Univ. Wisconsin, Aug. 1978.

Stability of the Demand Function for Money: An Unresolved Issue

By THOMAS F. CARGILL AND ROBERT A. MEYER*

The stability of the demand function for money has received extensive attention over the past two decades. However, there is no precise meaning of the term stability in the literature. The issue is most often discussed in reference to time-series estimates of the function and generally based on three characteristics: 1) The demand for money can be explained by a small set of variables as determined by various statistical tests; 2) the function does not exhibit marked shifts over time; 3) the function is capable of generating reasonable forecasts outside of the interval of estimation.

Stephen Goldfeld (1973) and John Boorman in exhaustive surveys conclude that relatively simple formulations of the demand for money yield stable short- and long-run functions. Despite some negative evidence (see William Poole), stability of the demand function has been fairly well accepted, at least up to the last few years (Goldfeld, 1976).

The overwhelming majority of evidence is based on time-series models using constant coefficient estimation procedures. Yet, arguments can be developed to show that estimating a demand function for money via constant coefficient methods amounts to misspecification. The time varying characteristics of the demand function should be explicitly recognized in the estimation procedure to properly investigate the stability issue.

This study is organized around two objectives: First, to use a theoretical model of risk preferences to develop the opportunity cost aspect of the demand function for money implying time varying coefficients and second, to provide within and outside sample comparisons of constant and variable coefficient estimates of various demand function specifications.

*Professors, University of Nevada-Reno, and University of California-Berkeley, respectively. We would like to thank B. Klein and J. Carlson for providing a portion of the data used in this study.

I. Theoretical Formulation

One needs to explictly develop an aggregate demand theory based on an uncertainty setting at the micro level where individual decision makers may have *heterogeneous risk preferences*. In any given period of time assume each decision maker, indexed by θ, makes consumption and investment (portfolio) decisions so as to maximize the expected utility of end of period wealth $W(\theta)$, i.e., end of period portfolio value, given market prices, current income $Y(\theta)$, and initial wealth $W_o(\theta)$:

$$(1)\quad W(\theta) = W_o(\theta) + Y(\theta) + R(\theta) - C(\theta)$$

where $R(\theta)$ is the current period return on the individual's portfolio and $C(\theta)$ is consumption. For small changes in wealth, John Pratt has shown that the exponential is the only continuous, monotonically increasing, and concave function with a constant (positive) absolute risk-aversion index; therefore, decisions for an individual of type θ are based on maximizing.

$$(2)\quad U[W(\theta)] = \alpha(\theta) - \gamma(\theta)e^{-r(\theta)W(\theta)}$$

where $r(\theta)$ is the index of absolute risk aversion and $-\infty < \alpha(\theta) < +\infty$ and $0 < \gamma(\theta) < +\infty$.

Assuming the yields on the n financial market instruments are approximately multivariate normal (for small changes), i.e., $p \sim N(p^*, \Sigma)$,

$$E[U[W(\theta)]] = \alpha(\theta) - \gamma(\theta) \operatorname{Exp}[-r(\theta)\{E[W(\theta)] - \frac{1}{2}r(\theta)\sigma_W^2\}] \tag{3}$$

Thus maximizing $U[W(\theta)]$ is equivalent to maximizing the certainty equivalent in the exponent in (3), that is, a mean-variance criterion which can be expressed as

$$\max_x U^*[W(\theta)] = A + p^{*\prime}x + r^*(\theta)x'\Sigma x \tag{4}$$

where $r^*(\theta) = r(\theta)/2$. The combination of the positive definiteness of Σ, the fact that $r^*(\theta) > 0$, and the quadratic character of (4) imply *unique derived demand functions*

$$x^*(p^*, \theta) = [r(\theta)]^{-1}\Sigma^{-1}p^* \tag{5}$$

Thus the derived demand function for each decision maker depends on three elements: 1) individual risk preferences; 2) expected market yields in period t; and 3) variances and covariances of portfolio yields in period t.

From (5) it is clear that a regression of $x^*(p^*, \theta)$ on p^* will yield coefficients which are a combination of both Σ^{-1} elements and $r(\theta)$ Such a regression is theoretically linear in p^* and could be statistically represented as

$$x_j^*(p^*, \theta) = \sum_{i=1}^{n} \beta_{ij}(\theta)p_i^* + v_j, \quad j = 1, \ldots, n \tag{6}$$

where $\beta_{ij}(\theta) = [r(\theta)]^{-1}\sigma^{ij}$ with σ^{ij} denoting the i, j element of Σ^{-1}.

Estimation of an *aggregate* demand function for any asset category requires summation over decision types θ, but there may be more than one individual with the same risk preferences. Let $h(\theta)$ denote the density function of risk preference types, hence the aggregate demand functions are

$$x_j^*(p^*) = \int_\theta x_j^*(p^*, \theta)h(\theta)\,d\theta \tag{7}$$

$$= \sum_{i=1}^{n} \{\sigma^{ij} \int_\theta \beta_{ij}(\theta)h(\theta)\,d\theta\}\, p_i^* + v_j$$

$$= \sum_{i=1}^{n} \beta_{ij}^* p_i^* + v_j$$

Estimated coefficients depend *explicitly* on the distribution of risk preferences at time t. Changes in the distribution of risk preferences (which could be called changes in expectations) will in general change system responses *even though objective financial market information may remain unchanged.*

II. Econometric Considerations

The above discussion indicates a need to employ time varying coefficient approaches. Two general methods are used here.

Systematic Coefficient Evolution Model: This method has been employed by the authors in a previous study and represents a generalization of many time varying estimation procedures. The method is based on representing one or more coefficients by a polynomial approximation in time (and a lag if dealing with distributed lag formulations).

State of the Economy Model: The state variable model employed by Meyer assumes that the coefficients of the demand function for money are a function of one or more variables that describe the "state of the economy." Expressing the coefficients in $Y = X\beta + u$ as a function of state variables,

$$\beta = Z\gamma \tag{8}$$

where Z is a k x l matrix of state variables and γ is the l x 1 vector of state variable coefficients. A linear transformation of γ^* yields estimates of the coefficients at each point in time. Stochastic variations of (8) are also possible.

III. In Sample Comparison Between Constant and Variable Coefficient Models

Both estimation procedures are used to estimate a demand function for money under a variety of specifications. Table 1 presents 24 specifications and the data sources. The 24 specifications are employed for annual data; however, in the case of quarterly data, 24 additional specifications are used by including a lagged dependent variable in specifications 1–24. The array of specifications examined spans a gamut from those frequently employed in the literature to those implied by the theoretical derivation presented above.

TABLE 1—SPECIFICATIONS OF THE DEMAND FUNCTION FOR MONEY AND DATA SOURCES

Specification	
1. $M_t = f(SR_t, Y_t)$	13. $M_t = f(LR_t, E_t, Y_t)$
2. $lnM_t = f(lnSR_t, lnY_t)$	14. $lnM_t = f(lnLR_t, lnE_t, lnY_t)$
3. $M_t/P_t = f(SR_t, Y_t/P_t)$	15. $M_t/P_t = f(LR_t, E_t, Y_t/P_t)$
4. $ln(M_t/P_t) = f(lnSR_t, ln(Y_t/P_t))$	16. $ln(M_t/P_t) = f(lnLR_t, lnE_t, ln(Y_t/P_t))$
5. $M_t = f(LR_t, Y_t)$	17. $M_t = f(SR_t, LR_t, E_t, Y_t)$
6. $lnM_t = f(lnLR_t, lnY_t)$	18. $lnM_t = f(lnSR_t, lnLR_t, lnE_t, lnY_t)$
7. $M_t/P_t = f(LR_t, Y_t/P_t)$	19. $M_t/P_t = f(SR_t, LR_t, E_t, Y_t/P_t)$
8. $ln(M_t/P_t) = f(lnLR_t, ln(Y_t/P_t))$	20. $ln(M_t/P_t) = f(lnSR_t, lnLR_t, lnE_t, ln(Y_t/P_t))$
9. $M_t = f(SR_t, E_t, Y_t)$	21. $M_t = f(SR_t, LR_t, Y_t)$
10. $lnM_t = f(lnSR_t, lnE_t, lnY_t)$	22. $lnM_t = f(lnSR_t, lnLR_t, lnY_t)$
11. $M_t/P_t = f(SR_t, E_t, Y_t/P_t)$	23. $M_t/P_t = f(SR_t, LR_t, Y_t/P_t)$
12. $ln(M_t/P_t) = f(lnSR_t, lnE_t, ln(Y_t/P_t))$	24. $ln(M_t/P_t) = f(lnSR_t, lnLR_t, ln(Y_t/P_t))$

Note: M = The money supply is defined as M_2 for annual data and M_1 for quarterly data.
Y = Net National Product is used for annual data and Gross National Product for quarterly data.
SR = The short-term interest rate is represented by the commercial paper rate for annual and quarterly data
LR = The long-term interest rate is represented by the corporate bond rate for annual and quarterly data.
E = For annual data, E is represented by Cowles Commission yield on common stock (1880–1928) and Moody's composite yield on common stock (1929–75). The dividend-price ratio is used for the quarterly data
P = Implicit price deflator for GNP is used for both annual and quarterly data.
Data obtained from U.S. Department of Commerce, *Long Term Economic Growth 1860–1970* and *Historical Statistics of the United States: Colonial Times to 1970*, Board of Governors *Econometric Model Data Base*, and Benjamin Klein.

Estimates based on annual data employ the period 1890–1970 while the period 1971–75 is used for forecast comparisons. Estimates based on the quarterly data employ the period 1954I–1974III and 1974IV–1975IV is used for forecast comparisons.

Systematic coefficient evolution estimates were obtained for each specification of the demand function for money using a third degree polynominal. The coefficients at each point in time often have the expected sign and significance and appear to be heavily time dependent. The results provide the basis for a general test of the nonconstant vs. constant coefficient model. Table 2 presents a summary of the test between constant and variable coefficient estimates of the demand for money. The constant coefficient model is rejected at the .05 level. While suggestive of significant time variation in the demand function for money, the results rely on polynominal approximations in time without any attempt to specify the cause of the time varying behavior.

The state variable model allows specification of the time varying behavior of the coefficients in the money demand as a function of one or more variables that represent the "state of the economy." In the case of the annual data, the inverse of the unemployment rate and an index of "economic discomfort" are used as state variables. The index of economic discomfort is defined as the sum of the unemployment and inflation rate and has been suggested as a general measure of economic confidence (see Michael Lovell). In the case of the quarterly data, the inverse of the unemployment rate, the index of economic discomfort, and a measure of unanticipated inflation (see J. Bisignano; John Carlson) are employed as state variables.

The F-statistics comparing the state variable with the constant coefficient model do not categorically reject the constant model using annual data. In the case of quarterly data specifications, less than half of the F-statistics failed to reject the constant coefficient model for each of the three state variables. Table 2 presents summary results.

IV. Forecast Comparisons

The results in the previous section suggests that constant coefficient estimates omit

TABLE 2—SUMMARY OF RESULTS COMPARING THE CONSTANT AND VARIABLE COEFFICIENT ESTIMATES[a]

	Number of Specifications	Number of F-Statistics Significant at .05 Level	Number of Lower Mean Squared Errors: Constant Coefficient Estimates	Number of Lower Mean Squared Errors: Variable Coefficient Estimates
Systematic Evolution Model[b]				
Data Type				
Annual (1890–1975)	24	24	6	18
Quarterly (1954I–1975IV)	48	48	13	35
State Variable Model				
Data Type				
Inverse of Unemployment				
Annual	24	10	10	14
Quarterly	48	34	24	24
Economic Discomfort Index				
Annual	24	14	17	7
Quarterly	48	34	20	28
Unanticipated Inflation				
Quarterly	48	48	23	25

[a]The F-Statistics and test of the constant versus the variable coefficient model are based on the period 1890–1970 and 1954I–1974III for the annual and quarterly data, respectively. The forecast comparisons are based on forecasts outside of the interval of estimation for the period 1971–75 and 1974IV–1975IV for the annual and quarterly data, respectively.

[b]The systematic evolution results are obtained for a third-degree polynomial.

important information since several varying coefficient estimation procedures significantly improved explanatory power within the sample. However, this represents only one aspect of the comparison between constant and varying coefficient estimates of the demand function. A central issue is whether improved forecasting characteristics can be achieved with time varying estimates.

Each of the estimated models was used to forecast over a five-period horizon outside of the interval of estimation. The mean squared error was used as the measure of relative predictive performance. Table 2 reports the summary results. The results can be summarized by the following points. 1) The polynomial estimates generally yielded lower mean squared error forecasts than the constant coefficients estimates for annual data specifications. In the case of the quarterly data specifications without lagged dependent variables, the polynomial method still yielded lower mean squared errors; however, when the specifications included the lagged dependent variable, only about half of the time varying specifications produced lower mean squared errors. 2) The state variable model using the inverse of the unemployment rate yielded lower errors in 14 out of 24 specifications for the annual data; however, using the economic discomfort index as a state variable yielded lower forecast errors in only 7 out of the 24 specifications. 3) For quarterly data, the state variable model generally performed worse than the constant coefficient model without a lagged dependent variable. 4) The state variable models generally forecast better than the constant coefficient model when the specifications include a lagged dependent variable, except when the inverse of the unemployment rate was used as a state variable.

The evidence based on forecasting comparisons does not categorically reject the constant coefficient estimates in every case; however a number of specifications of the

demand for money yielded lower forecast errors when the coefficients were allowed to vary over time either in some systematic manner or when some variable(s) describing the state of the economy was employed.

A general comment should be made about the forecast comparisons. The results for many specifications indicate improved forecasts over constant coefficient models; however, the improvement is not very dramatic and one may very well question whether the additional effort and information needed to estimate variable coefficient models is advisable. While the variable coefficient forecasts often yield lower mean squared errors they also require additional information and estimated parameters over constant coefficient estimates. The question cannot be answered at this stage. The present study is primarily exploratory and is only suggestive that variability should be taken into account in the estimation of demand functions for money.

V. Conclusion

Theoretical arguments imply a time-varying response of the demand for money to income and opportunity cost changes, thus the application of constant coefficient estimation procedures will yield biased tests of stability. Despite the large body of empirical evidence supporting the existence of a stable function over short and long periods of time, the issue still appears to remain open. The evidence presented in this study is based on a wide variety of specifications and both annual and quarterly data.[1] Simple *OLS* estimates yield results with high R^2 values and coefficients with correct signs and significance. To all appearances the results support the existence of a stable function, yet when the same functional forms are estimated by procedures that allow for nonconstant coefficients the overwhelming majority of cases yield a rejection of the hypothesis of constant coefficients. Forecast comparisons also suggest that varying coefficient models outperform constant coefficient models for several specifications, thus there does not appear to be firm support for the conclusion that the demand function for money is temporally stable.

[1]The estimates in this paper are based on *OLS* estimates of the basic equations needed to derive time varying estimates of the demand function for money. Problems of serial correlation and simultaneity have not been addressed in the context of these estimation procedures given the exploratory nature of the study. Future efforts will include a smaller set of specifications that will allow consideration of estimation and interpretation issues not addressed in this study.

REFERENCES

J. Bisignano, "Savings, Money Demand and the Inflation/Unemployment Tradeoff," *Fed. Reserve Bank San Francisco Rev.*, Summer 1977, 6–20.

J. T. Boorman, "The Evidence on the Demand for Money: Theoretical Formulations and Empirical Results," in Thomas M. Havrilesky and John T. Boorman, eds., *Current Issues in Monetary Theory and Policy*, 1976, 315–60.

J. Carlson, "A Study of Price Forecasts," *Annals Econ. Soc. Measure.*, Winter 1977, *6*, 27–56.

T. F. Cargill and R. A. Meyer, "The Time Varying Response of Income to Changes in Monetary and Fiscal Policy," *Rev. Econ Statist.*, Feb. 1978, *60*, 1–7.

S. M. Goldfeld, "The Demand for Money Revisited," *Brookings Papers*, Washington 1973, *3*, 577–638.

———, "The Case of the Missing Money," *Brookings Papers*, Washington 1976, *3*, 683–730.

B. Klein, "Income Velocity, Interest Rates, and the Money Supply Multiplier: A Reinterpretation of the Long-Term Evidence," *J. Money, Credit, Banking*, May 1973, *5*, 656–68.

M. C. Lovell, "Why Was the Consumer Feeling So Sad?," *Brookings Papers*, Washington 1975, *2*, 473–479.

R. A. Meyer, "Structural Stability and Policy Analysis with Macrodynamic Models," *Southern Econ. J.*, Oct. 1977, *44*, 249–60.

W. Poole, "Whither Money Demand?," *Brookings Papers*, Washington 1970, *3*, 485–500.

J. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

Board of Governors of the Federal Reserve System, *Quarterly Econometric Model Data Directory and Listing.*

U.S. Department of Commerce, *Long Term Economic Growth 1860–1970*, Washington June 1973.

———, *Historical Statistics of the United States: Colonial Times to 1970*, Washington 1975.

Structural and Technological Change in Money Demand

By CHARLES LIEBERMAN*

Although demand for money equations have traditionally been regarded as exceptionally stable, the literature abounds with empirical evidence of structural shifts and apparent trends in some of the coefficients over time. Subperiod estimates of long-term studies often yield different coefficients or suggest secular trends in the coefficients. Other studies report structural shifts or variables which are significant only during particular periods. And while numerous other instabilities have vanished along with revisions in the data, the 1974 instability is too large to be so obliging. Thus, the latest apparent instability is hardly without precedent.

One possible source of these instabilities is the omission from the estimated equations of a measure of technological change which reduces over time the real cost of transactions in the management of money balances. The real cost of transactions plays a substantial role in explaining money holdings in the standard inventory theoretic models of money demand. The omission of a key explanatory variable from the estimation produces a misspecified equation which may result in biased coefficient estimates. While the sources of the most recent instability in money demand equations may be due to a structural shift, as suggested in several studies, it may also be due in part to an omitted technological change measure.

This paper tests the usefulness of some technological change proxies and examines the forecasting ability of these modified models over the difficult post-1974 period. A Shiller lag specification is employed to obtain a better specification of the dynamic properties of money demand. Although the technological change proxies do improve the equations, the findings support the existence of another structural shift in money demand. The results nevertheless provide some hope that a fairly standard money demand equation, with only minor adjustments, may once again be capable of providing satisfactory forecasts, at least for the present.

I

Standard Keynesian money demand models employ income as a proxy for transactions and are linear in the *logs* of the variables. The stock adjustment model appears to be the procedure of choice in virtually all empirical studies of the dynamic behavior of money demand. (Stephen Goldfeld observed that Almon lags provide results similar to those produced by Koyck lags.) The analysis reported here employs the lag procedure introduced by Robert Shiller instead of the stock adjustment model because the estimated speeds of adjustment provided by the Koyck lag have always been regarded as surprisingly long. Long lags are consistent with the theoretical biases introduced into estimates by the combination of a lagged dependent variable and serial correlation. Instead of forcing the coefficients to lie along a polynomial (as with the Almon lag) or decay exponentially (as with the Koyck lag), the Shiller lag imposes a degree of smoothness on the coefficients of the lagged exogenous variables. Since the Shiller lag procedure includes only contemporaneous and lagged values of the exogenous variables, the biases introduced by a lagged dependent variable are avoided. And unlike the Almon lag (which is a special case of the Shiller lag), the Shiller lag method fits tails easily, a charac-

*Visiting associate professor of economics, Northwestern University. This paper was written while I was on the faculty of the University of Maryland. Thanks go to Leslie Czubek for research assistance and Thomas Mayer, Eileen Mauskopf, Richard Porter and Martin Regalia for comments, but I retain responsibility for any errors.

teristic which assumes critical importance when the lag lengths are quite short.

The basic model employed in the estimated equations is

$$(1)\quad \ln M_t = \alpha_o + \sum_{i=0}^{k_1} \beta_i \ln Y_{t-i} + \sum_{i=0}^{k_2} \delta_i \ln R_{t-i} + \sum_{i=0}^{k_3} \gamma_i \ln W_{t-i} + \lambda t$$

where M is the real money stock (M_1), Y represents transactions as modeled by real *GNP*, R represents one or more interest rates, W is real household wealth (M.I.T.-Pennsylvania-Social Science Research Council-*MPS* definition), t represents the technological change proxy, and the lags are estimated using the Shiller technique. Except for the money data which are quarterly averages, the data are from *MPS* data bank. By including wealth and income in the equation, the explanatory power of the transactions and asset theories can be compared unambiguously. When a Koyck lag is employed, it is impossible to distinguish the transactions model from an asset model which employs permanent income.

Measuring the impact of technological change on money demand is quite difficult. Conceptually, it would be most desirable to model explicitly each source of improvement in cash management. For example, reductions in money holdings due to increased use of credit could be modeled by the number of credit cards outstanding or the volume of credit card use. (This method has previously been employed with little success.) Unfortunately for empirical research, innovations in cash management include dozens of new activities, instruments, institutions, and practices for which data are often sparse or nonexistent. Credit cards, lock boxes, zero balance accounts, telephone transfers, automatic billpayer accounts, direct deposit of payrolls and receipts, cash disbursal accounts, cyclical billing, and wire transfers have all increased the velocity of money. And even if it were feasible, modeling each source of innovative change in cash management would quickly exhaust the degrees of freedom available.

Omitting any measure of technological change is hardly costless, however. Theoretically, real transactions costs belong in the equation to be estimated. And if technological change is overlooked, the coefficient estimates will be biased.[1] As a result, even crude proxies for the omitted variable, technological change, ought to be considered.

One proxy for technological change which is commonly used elsewhere is a time trend. The inclusion of a time trend in the estimation merely assumes that the implementation of new technologies or practices reduces money demand smoothly over time. Although the introduction of new technologies or cash management methods may be discrete or lumpy rather than continuous, the implementation of these new methods throughout the economy will be distributed with a (possibly long) lag thereby smoothing out the impact on money balances. A time trend would measure the mean rate at which new cash management techniques reduce money balances, *ceteris paribus*, as shown in my 1977 paper.

Unlike the time trend which assumes that technological change is exogenous to the financial sector, a second proxy, introduced by Perry Quick and John Paulus, assumes that improvements in cash management are largely induced by rising opportunity costs. Quick and Paulus employ a past-peak interest rate proxy where the "peak" decays over time until another rise in interest rates creates a new peak. When interest rates rise above the previous threshold, new peaks are set which induce improvement in cash management. If rates decline, the threshold is permitted to decay gradually as some expensive cash management techniques are scrapped and as exogenous technological change requires a lower threshold to spawn a new wave of economizing of money balances.

A third possible proxy for technological change in money demand equations is the past-peak per capita level of real *GNP*, a

[1] An omitted relevant variable biases estimated coefficients of the included variables if any of the included variables are correlated with the omitted variable. Since real *GNP*, the usual transactions measure, includes the effects of economy wide technological change (and is also a growth variable), the necessary condition for bias in the coefficients is likely to be satisfied.

TABLE 1—A QUARTERLY DYNAMIC MONEY DEMAND EQUATION

	period t	t − 1	t − 2	t − 3	t − 4	Σ
GNP	.118	.161	.157	.103	.069	.609
	(2.15)	(3.95)	(3.75)	(2.52)	(1.24)	(2.60)
RCP	−.015	−.018	−.016	−.013	.001	−.061
	(2.95)	(3.34)	(2.79)	(2.35)	(.25)	(2.26)
	constant = 1.47 (2.02)		time = −.00313 (2.88)		wealth = .127 (3.54)	
R^2 = .996		*S.E.* = .0043		p = .95		*D.W.* = 1.41

Note: The numbers in parenthese are t-ratios. All of the variables enter linearly in logs except for the time trend. *RCP* is the commercial paper rate.

measure which has been employed in some versions of the Federal Reserve's money demand equation.[2] Like the time trend measure, this variable may be thought of as also assuming that technological change is exogenous to the money market so that the financial sector adopts innovations introduced elsewhere in the economy.

II

Table 1 and Figure 1 report the results of the estimation of equation (1) where a time trend is used as the proxy for technological change and the distributed lags are estimated using the Shiller method (with no end-point restriction). The degree of tightness of the prior, the k value, is .05 but the results are not especially sensitive to the k value employed. The equation is adjusted for serial correlation using a Hildreath-Liu search routine.

The results indicate that money demand is largely a function of a distributed lag of income and transactions while wealth and asset motives play, at most, a minor role. The income effect builds to a peak before trailing off, a feature which Koyck lags cannot produce. The sum of the lagged income effects is .609 and is statistically different from zero and from unity at the 90 percent level. The finding relative to unity suggests substantial economies of scale, as indicated by the inventory theoretic approach to money demand. The magnitudes of the effects reported here are also in excellent agreement with Goldfeld who used quarterly data and a Koyck lag and with my 1977 paper using annual data and bank debits as a measure of transactions.

The wealth coefficient is highly significant and of the theoretically correct sign but quite small. Unlike the income or interest rate effects, the magnitude of the estimated wealth effect varied inversely with the lag length. With the exception of the contemporaneous term, distributed lag coefficients were generally close to zero. The contemporaneous term, however, tended to remain sizeable and significant, hence its inclusion in the equation presented. Even so, the estimated wealth elasticity is well below unity, which indicates that money is a necessity rather than a luxury asset. This finding is consistent with the cross-section evidence and the recent time-series evidence of the author. When wealth was excluded entirely from the equation, the income coefficients were increased modestly while the fit and the out-of-sample forecasts were largely unaffected.

The interest rate coefficients indicate that the effect on money demand builds to a peak which is reached within two quarters before it tails off to zero. Longer lags produce only additional small and insignificant coefficients while the sum of the coefficients is virtually unchanged. The sum of the coefficients at −.061 suggests that money is rather insensitive to changes in market interest rates, a finding shared with Goldfeld and which is examined by the author (forthcoming). Numerous other interest rates and combina-

[2]This measure does not distinguish between per capita output growth due to technological advance and per capita output growth due to capital investment.

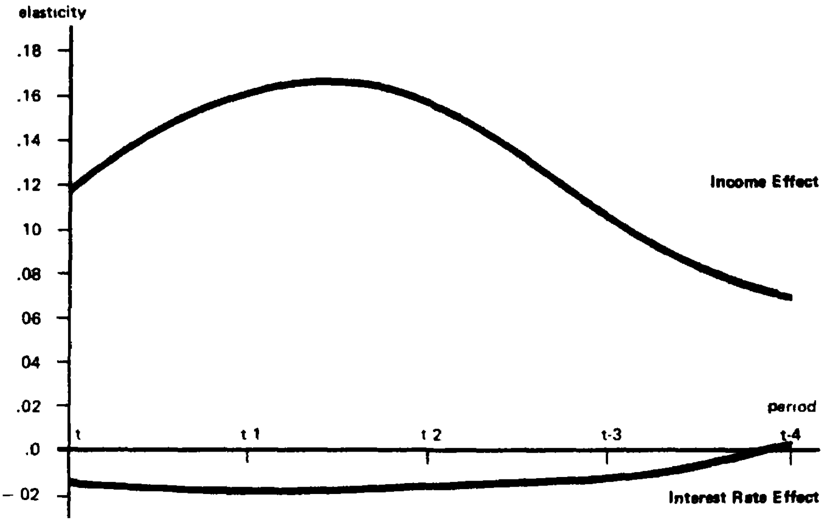


FIGURE 1. A QUARTERLY DYNAMIC MONEY DEMAND EQUATION
Sample Period 1952II–1974II

tions of interest rates were tested. The commercial paper rate fit and forecast about as well as any other combination and better than most. Michael Hamburger's combination of the dividend rate, bond rate, and passbook deposit rate did not fare as well as the commercial paper rate.

The coefficient of time, the proxy for technological change, indicates an effect of technological change on money demand of about 1.2 percent per year. This result is very close to the effect of 1.3 percent reported by the author (1977) for annual data and which employs bank debits instead of income as a measure of transactions. The remaining two proxies for technological change, the past-peak interest rate and past-peak per capita real income, were statistically insignificant although the past-peak interest rate measure provided coefficients of the theoretically correct sign. Thus, the time trend emerged as the most effective of the technological change proxies examined.

The dynamic properties of the estimated equations are more reasonable than the results suggested by equations which include a lagged dependent variable. In particular, the estimated speeds of adjustment of equations which include lagged dependent variables are often extremely slow, as little as 2 to 3 percent adjustment per quarter, and rarely as rapid as 70 percent per year. These findings seem inconsistent with the widely held view that financial markets clear quickly.

The coefficient estimates suggested by the Shiller lag procedure reach complete adjustment far more rapidly than other estimation methods. Money demand adjustment with respect to income is almost 50 percent complete within two quarters and 90 percent complete within one year. The response with respect to the commercial paper rate is complete within one year while adjustment is complete with respect to wealth within one quarter. Moreover, the lag coefficients do tail off to zero even though no end-point restriction is imposed.

III

Despite the good fit and the magnitudes of the coefficients which are well within the bounds suggested by theory, the forecasting quality of this equation deteriorates markedly after the second quarter of 1974. This is the same problem period for equations which employ the traditional Koyck lag method and which exclude a proxy for technological

change. The root-mean squared error from the third quarter of 1974 through the fourth quarter 1977 is $19.56 billion but only $11 billion if the actual errors of the extrapolation are used. While these errors are smaller than the errors generated by the Federal Reserve's *MPS* equation and no larger than any other formulation, they are still quite substantial. Moreover, none of the proxies for technological change eliminated the substantial forecasting errors for this period. Furthermore, regardless of the independent variables employed in the estimation (including Hamburger's variables which were reported to have provided much reduced errors), the forecast errors were large and highly correlated, suggesting a common cause.

Even so, the errors in the forecast period peak during 1976 and remain virtually static within a range of −$25 to −$28 billion through all of 1977. Thus, whatever the source of the shift in money demand—and none of the technological change proxies accounts for it—the structural shift seems to have stabilized, at least for about six quarters. Rather than a continuous and continuing shortfall in money demand, the mystery now appears to be a once and for all deviation of money demand behavior from a norm which had persisted for some time to what appears potentially a new norm.

It is rather simple to account for the timing and at least some of the magnitude of this apparent structural shift by several key events which occurred in 1974. These events include the introduction of *NOW* accounts throughout New England, the introduction and rapid growth of money market mutual funds, the issue of share drafts by federal credit unions, the permission granted banks to offer savings accounts to state and local governments and to corporations, the development of the repurchase agreement (*RP*) market, and the development and sale of cash management services as well as the additional recent developments cited earlier.

Paulus and Stephen Axilrod have very conservatively estimated that the factors cited above (excluding *RP*'s and cash management services) could account for about $5.5 billion in the shortfall of actual from projected growth in money demand by the third quarter of 1976. A less conservative but reasonable evaluation of their figures could easily double the estimated impact of these events. Since the errors generated by this equation have stabilized within the $25 to $28 billion range, the less conservative impact of the factors cited could easily account for over one third of the shortfall in demand. Moreover, several other factors, principally the *RP* market and the marketing of cash management services could easily account for the remainder.[3]

If these events are in fact behind the shortfall in the growth of money, we would expect that once these new instruments or practices had been incorporated fully into the public's behavior patterns, that the deviation of actual from forecasted money holdings would stabilize, as has indeed occurred. And given the diversity and significance of the new instruments and institutions, it is hardly surprising that it has taken about two years, from mid-1974 to mid-1976, for the economy to have adjusted its money holdings to these new arrangements.

The arguments presented suggest that the money supply, having absorbed the once and for all introduction of new institutional arrangements, may well revert to its "normal" behavior. If this analysis is correct and barring any significant and dramatic new institutional innovations in the money market, traditional equations, which are adjusted for a once and for all structural change in the 1974–76 period, should be able to forecast with errors no larger than those produced prior to 1974. But while this conclusion appears comforting, given the variety and fundamental nature of the changes in banking practices presently being considered by the Congress and the Federal Reserve, it would seem to be merely a matter of time before substantial additional structural changes are permitted which will once again allow the

[3]Richard Porter and Eileen Mauskopf believe that the marketing of cash management services accounts for much of the reduction in money balances. Elsewhere in this session, Gillian Garcia and Simon Pak attribute the overestimates to the development of the *RP* market. Both phenomena undoubtedly contributed substantially to reducing money balances.

public to economize quantumly rather than continuously on its holdings of money balances.

REFERENCES

S. M. **Goldfeld**, "The Demand for Money, Revisited," *Brookings Papers*, Washington 1973, *2*, 577–638.

M. J. **Hamburger**, "Behavior of the Money Stock: Is there a Puzzle?," *J. Monet. Econ.*, July 1977, *3*, 265–88.

C. **Lieberman**, "The Transactions Demand for Money and Technological Change," *Rev. Econ. Statist.*, Aug. 1977, *59*, 307–17.

_____, "A Transactions vs. Asset Demand Approach to the Empirical Definition of Money," *Econ. Inquiry*, forthcoming.

J. **Paulus** and S. H. **Axilrod**, "Recent Regulatory Changes and Financial Innovations Affecting the Growth of the Monetary Aggregates," unpublished manuscript, Fed. Res. Board 1976.

R. D. **Porter** and E. **Mauskopf**, "Some Notes on the Apparent Shift in the Demand for Demand Deposits Function," unpublished paper, Fed. Res. Board 1978.

P. D. **Quick** and J. D. **Paulus**, "Financial Innovation and the Transactions Demand for Money," unpublished manuscript, Fed. Res. Board.

R. J. **Shiller**, "A Distributed Lag Estimator Derived from Smoothness Priors," *Econometrica*, July 1973, *41*, 775–88.

Some Clues in the Case of the Missing Money

By GILLIAN GARCIA AND SIMON PAK*

Stephen Goldfeld revealed that the received demand functions for demand deposits and M_1, seriously overpredicted demand during the recent period.

$$(1)\quad \ln m_{it} = a_{i0} + a_{i1} \ln y_t + a_{i2} \ln RTD_t + a_{i3} \ln RCP_t + a_{i4} \ln m_{it-1} + e_{it} \qquad i = 1 \ldots 3$$

which is Goldfeld's (logarithmic) demand for real stocks of $M_1 (i = 1)$, demand deposits $(i = 2)$, and currency $(i = 3)$ as a function of real GNP (y_t), of the interest rates on time deposits (RTD_t) and on commercial paper (RCP_t), and of the lagged money stock. In his estimation Goldfeld used data from the Board of Governors of the Federal Reserve System for the period 1952.2–1973.4. His results are presented in Table 1, where they are compared to our replication of his experiment using data from the same source for the same period. The discrepancies in parameter estimates may be attributed to data revisions.

However, when the equations (1) are reestimated over the period 1952.2 through 1976.2, the results (also presented in Table 1) are unacceptable in several regards. In the equation for M_1, the parameter estimates for *GNP* and *RTD* are not significant, that for the interest rate on time deposits has the wrong sign, and the speed of adjustment is extremely slow. The decomposition of M_1 into currency and demand deposits reveals that the currency equation continues to explain demand well over the longer period and confirms Goldfeld's assertion that the demand for demand deposits is the source of the money demand problem. During the longer period the equation for demand deposits is totally unacceptable. The coefficient of *GNP* is insignificant, the coefficient of *RTD* has the wrong sign, and the coefficient of the lagged endogenous variable is greater than one. Further, the income elasticities of both M_1 and demand deposits are significantly smaller by an order of magnitude in the longer period regressions than they are in the shorter period. The income elasticity of currency demand is, however, unchanged in the two estimations.

It is surprising that the addition of ten new observations to the money demand function's estimation period invalidates the previously accepted theory. It would appear that a structural change (confirmed by an F-test) has occurred in the demand deposit function. Being dissatisfied with the predictive performance of this function, Goldfeld estimated several respecifications. These were also examined by the authors and found to be unsatisfactory.

I. The Scene of the Crime

Reestimation of Goldfeld's respecifications over the longer period confirmed that the failure was more pronounced in the business than in the household sector's equations. This inference is supported by the results presented in Table 1, for the real per capita, nominal adjustment respecification of the money demand function (see Goldfeld, 1976, pp. 691–92) which was reestimated using data on sectoral deposit holdings for the period 1970.2 through 1976.2. The data were obtained from the Federal Reserve's Survey of Demand Deposit Ownership. The household sector's equation remains valid (although the coefficient of *RTD* is positive), but the business sector's is not. Again, its principal problems

*University of California-Berkeley. Research for this paper was supported by a grant from the Dean Witter Foundation administered through the Institute of Business and Economic Research. We thank Raymond Lombra, Thomas Mayer, James Pierce, and David Pyle for comments and suggestions and Charles Salazar for assistance under the university's Professional Development Program, but retain responsibility for any errors.

TABLE 1—CONVENTIONAL MONEY DEMAND EQUATION FOR ALTERNATIVE MEASURES OF MONEY BALANCES 1952.2–73.4 and 1952.2–76.4

			Coefficient[a]				Summary Statistic		
				Interest Rate					
Dependent Variable	Time Period	Author	Income	Time Deposits	Commercial Paper	Money Lagged	R^2	Standard Error	ρ
Currency plus Demand	1952.2–73.4	Goldfeld	0 179	−0.042	−0.018	0 676	0.995	0.0042	0.35
Deposits, M_1			(5.4)	(4.0)	(6.5)	(10.0)			
Currency plus Demand	1952 2–73.4	G-P	0 223	−0.052	−0.023	0.600	0.993	0.0047	0.29
Deposits, M_1			(6 1)	(4 6)	(7 8)	(8.2)			
Currency plus Demand	1952 2–76.2	G-P	0.029	0 006	−0.022	0.977	0.988	0.0063	0.29
Deposits, M_1			(1.3)	(0 8)	(5.8)	(20.7)			
Money plus *IAFs*	1952 2–76 2	G-P	.106	−.019	−.020	.825	0 988	0072	0 47
(method 1)			(3.6)	(1 6)	(4 1)	(13.5)			
Money plus *IAFs*	1952 2–76 2	G-P	176	− 039	− 020	769	993	0071	.36
(method 2)			(5.1)	(3.3)	(4.6)	(14 3)			
Demand Deposits	1952.2-73.4	Goldfeld	0.158	−0 034	−0 020	0 661	0.991	0.0048	0 36
			(5.4)	(3 6)	(6.3)	(8 9)			
Demand Deposits	1952 2-73 4	G-P	0.203	−0 044	−0 026	0 570	0.987	0.0056	0.26
			(6.4)	(4 5)	(7 7)	(7 5)			
Demand Deposits	1952.2–76 2	G-P	0.010	0.013	−0.025	1 035	0 976	0.0074	25
			(0 6)	(1 9)	(6 1)	(26.76)			
Demand Deposits	1970 2–76.2	G-P	0 246	0.309		0.704	793	0 0205	−0 21
Household			(1 9)	(3 0)		(6.6)			
Demand Deposits	1970 2-76 2	G-P	0.060		−0 490	1 129	965	0085	−0 09
Business			(1 5)		(6.6)	(23 9)			
Demand Deposits plus *IAFs*	1952 2–76.2	G-P	.081	− 010	− 023	875	982	0086	.40
(method 1)			(3 0)	(0 9)	(4 3)	(15 4)			
Demand Deposits plus *IAFs*	1952 2–76 2	G-P	154	− 032	−.023	794	987	0090	35
(method 2)			(4 5)	(2 5)	(4 0)	(13 8)			
Currency	1952 2–73 4	Goldfeld	0.H7	−0 026	−0 007	0 863	0 998	0.0043	0 62
			(3 5)	(2 3)	(1.9)	(21 6)			
Currency	1952 2–73 4	G-P	0 129	−0 027	−0 010	0 853	0 998	0 0040	0.62
			(4 1)	(2.6)	(2.9)	(21 9)			
Currency	1952 2–76.2	G-P	0 129	−0 026	−0 013	0 864	0 998	0 0043	0 52
			(4 2)	(2 5)	(4.1)	(24.3)			

Source Basic data are from the Board of Governors of the Federal Reserve System *MPS* model

[a]All equations contain an intercept, which is not reported, and all variables enter logarithmically The numbers in parentheses are t-ratios In the Garcia-Pak equations the Cochrane-Orcutt method of estimation was used

are the coefficient of the lagged endogenous variable, which is greater than unity, and that of income, which is small and insignificant. It is not surprising, therefore, that those institutional innovations, such as negotiable order of withdrawal (*NOW*) accounts, money market mutual funds, credit cards, savings deposits of business and state and local governments, and checking accounts at mutual savings banks, which have been credited with the M_1 reduction are in fact unable to account for the bulk of the missing money. These innovations pertain mostly to the household sector, whereas the structural shift appears to be located primarily in the business sector.

Further, the conceptualization of these innovations as causing a reduction in money demand may be misleading if Will Mason's definition of money as "a direct and general claim on other assets exercisable on demand" is accepted, for under it some of the technological innovations should be classified as money not as substitutes for it. This is particularly true of a little known development in the federal funds market which we discuss below.

II. Immediately Available Funds Transactions

The prime suspect is the growth of immediately available funds (*IAFs*) transactions in the federal funds market. The practice has

proliferated during the 1970's for large depositors to sell overnight to their bank their end-of-day demand deposits. Next morning the bank returns the deposits for the corporation (or state and local government or other large depositor) for use during the day. Such a transaction, often secured by a repurchase agreement (*RP*), allows the depositor in effect to earn interest at, or a little below, the federal funds rate, on what remains as a transactions balance during the business day.

From the bank's point of view the advantage of an *RP* is that it substitutes in its balance sheet a liability to federal funds purchased in place of that to demand deposits. Since the revisions of Regulation D in 1969 and 1970, this transfer has allowed banks to reduce the value of required reserves because federal funds purchases were then declared nonreservable. To the extent that banks can regularly repeat such federal funds purchases with whoever is holding surplus funds, they allow banks to systematically reduce reserve pressure.

It seems likely, therefore, that the proliferation of *IAF*s transactions has reduced the demand for measured deposits, especially those of corporations, thrift institutions, state and local governments, government agencies, and dealers and brokers in government securities. It is unfortunate, therefore, that data on *IAF*s are not readily accessible, in published form. Their aggregate value is, in principle, measurable however. In this paper attention will be confined to two measures, one the smallest of the known estimates and the other the largest. The first method subtracts the value of all federal funds (including *RP*s) sold by commercial banks from those purchased by commercial banks, as recorded in the Federal Deposit Insurance Corporation's call reports. This gives a figure (recognized to be an underestimate because of "window dressing") for net federal funds purchased from nonbanks. The second method takes estimates of "Federal Funds Sold and Security *RP*s" from the Flow of Funds Accounts and adds to them "Float on Commercial Bank Interbank Loans." The reason for the addition is that the first series is based on sectoral balance sheets which do not always reveal *RP* activity, so that it underestimates the true value, as confirmed by the banking sector records.

Our suspicions were given some confirmation by these estimates of the aggregate value of *IAF*s for they appear large enough to be able to make a significant contribution to explaining the large dynamic simulation error in Goldfeld's forecasts. For example, the lowest *IAF*s estimates, those of net federal funds purchases, showed *IAF*s rising from \$1.7 billion in mid-1968 to \$24.6 billion by June 1976, while the largest, the Flow of Funds estimates, have them rising from \$2.9 billion to \$36.5 billion over the same period. The 1976 estimates span the current dollar value of \$29.6 billion for Goldfeld's dynamic forecast error in June 1976.

III. Are *IAF*s Money?

It seems possible, therefore, that *IAF*s transactions have caused a substantial reduction in deposit demand, and, in particular, in the business and government sectors' demands. Further, some *IAF*s may conceptually be regarded as money. Deposits are measured at the close of bank business, but any transformation of daytime deposits at the end of the business day which restores the deposit for use next day should conceptually be regarded as money. For it is immediately available as a claim against other assets during the business day which is the only time such a claim can be exercised.

Some *IAF*s, however, have maturity beyond one day and some are contracted before the close of bank business. Conceptually in an uncertain world these longer transactions curtail the seller's transactions balances and should properly be regarded as the replacement of money by a short-term asset. Consequently, aggregate *IAF*s data overstate the value of non-M_1 transactions balances. Federal Reserve Survey data for Decembers 1974 and 1977 show 79.3 percent and 77.1 percent, respectively, of gross nonreservable borrowings in immediately available funds were of one day or continuing contract duration, but data on the timing of

contract commitments are not publicly available at present. (See the *Federal Reserve Bulletin*.)

Although it is recognized that some *IAF*s are money and some are money substitutes, given the inability to distinguish them, monetary economists are left with the unsatisfactory alternatives of treating all as money or all as nonmoney. At present not only are *IAF*s regarded as nonmoney, but also they are excluded from the broader money stocks $M_2 \ldots M_5$, although they appear to be more liquid substitutes for money than time or savings deposits, or any other near money.

IV. The Evidence

The consequences of adhering to the current definition of M_1 have already been demonstrated: Goldfeld showed that the dynamic forecasts are biased with large and increasing errors and this paper has illuminated the unsatisfactory equation estimates. We have also shown in our 1978 paper that such mismeasurement leads to biased and inconsistent parameter estimates in the money demand function and have simulated the extent of the problem. In this situation several courses of action remain. Some economists have concluded that the money demand function, so fundamental to macro-economic theory and policy, is no longer stable. Alternatively, time and/or dummy variables have been introduced into the equation to restore the parameter estimates to conformance with prior expectations and to improve the forecast performance. Further, different interest rates have been utilized, *RP*s have been tried as an explanatory variable in the money demand functions, and variable coefficient estimates attempted.

This paper investigates another approach: treating *IAF*s as money. Table 1 shows the results for deposits and M_1 of reestimating equation (1) when two alternative estimates of *IAF*s are included in the money stocks. The coefficient estimates conform to prior expectations and to pre-1974 estimates in terms of signs and relative sizes.

Further, as Figure 1 demonstrates, the ability to predict M_1 is enhanced by basing the forecasts on an equation which includes *IAF*s in the dependent variable. The forecasts from money stock + *IAF* equations estimated over the period 1952.2 through 1967.4 (before the *IAF*s explosion occurred) are compared to forecasts from the conventional function. The root-mean square errors are lower ($9.96 billion or 4.18 percent of the mean real money stock compared to $13.22 billion or 5.54 percent for the period beginning in 1968.1 and $5.05 billion or 2.17 percent compared to

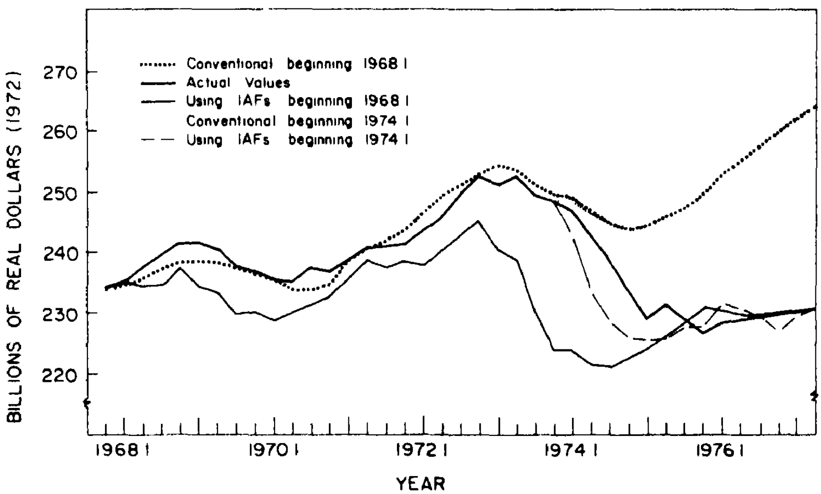


FIGURE 1. FORECASTS OF MONEY DEMAND

$21.57 billion or 9.28 percent for the Goldfeld period beginning in 1974) and the systematic bias problem is absent for the augmented money stock equation. However the forecasts for the years 1973 and 1974 remain unsatisfactory and suggest that the conventional augmented money demand specification needs still further refinement before it can explain money demand in that inflationary period.

V. The Close of the Case

The clues presented in this paper have suggested that much of the missing money may be located in the federal funds market. Other, smaller parts are to be found among the other technological developments in the money market. As many of these pertain principally to the household sectors, its money demand function problems may be expected to be solved similarly.

The evidence in this paper cannot be decisive in the conceptual issue of whether *IAF*s are money or not, but it does provide some support for the argument that on both practical and conceptual grounds, *IAF*s, if not money, are closer substitutes to it than other near monies and that they help explain the postrecession problems in estimates of the *U.S.* money demand function. As a consequece, timely and accurate data on *IAF*s has become necessary to understanding and predicting money demand. This should permit the incorporation of *IAF* behavior into the macro-economic models of the *U.S.* economy.

REFERENCES

G. Garcia and S. Pak, "Bias and Inconsistency in the Estimation of the Money Demand Function," mimeo., 1978.

S. M. Goldfeld, "The Case of the Missing Money," *Brookings Papers*, Washington 1976, *3*, 683–730.

W. E. Mason, "Can the Concept of Money be Saved?," mimeo., 1978.

Federal Reserve Board, "*MPS* Model Data Base."

Fed. Res. Bull., "Repurchase Agreements and Federal Funds," May 1978, *64*, 353–60.

Strategic Entry Deterrence

By STEVEN C. SALOP*

In analyzing deterrence of large-scale entry, two classes of entry barriers may be distinguished. An *innocent* entry barrier is unintentionally erected as a side effect of innocent profit maximization. In contrast, a *strategic* entry barrier is purposely erected to reduce the possibility of entry. Two types of innocent barriers may also be distinguished. A ***postentry absolute advantage*** has the property that, if entry did occur, the established firm would be at a profit advantage over the entrant. Examples are superior technology or product design, patents, and lower input prices. A ***preentry asymmetry advantage*** arises from the fundamental preentry asymmetry between established firm and potential entrant. Before the entrant makes his entry decision, the established firm has already committed resources. This prior existence gives first-move advantages. The preentry asymmetry is independent of symmetry or asymmetry in the rules (the equilibrium concept) of the postentry game that might ensue; even if the postentry game will be played according to Nash-Cournot or entrant-as-leader rules, the preentry leadership role always lies with the established firm.

For example, consider the innocent barrier due to scale economies. The entrant should ignore preentry price and profit levels, but attempt to infer the postentry equilibrium price and profit levels. If the entrant's expected profits are negative, he is deterred; the no-entry profits accrue to the already established firm rather than the equally efficient entrant. Even a more efficient entrant may be deterred by an established firm who has sunk sufficient costs to make his own exit uneconomical, and hence, entry mutually destructive.

This fundamental asymmentry in the preentry game provides the foundation for the theory of strategic barriers against equally efficient potential entrants. By making binding commitments and communicating them during the preentry period, a strategically minded established firm is able to exploit its leadership role. If the commitments imply negative profits to the entrant in the postentry game, then entry will be successfully deterred. As Thomas Schelling states, "the essence of these tactics is some voluntary but irreversible sacrifice of freedom of choice. They rest on the paradox that the power to constrain an adversary may depend on the power to bind oneself" (p. 22).

In the following sections, some examples of this view of strategic entry deterrence are presented. Section I focuses on perfect information, while Section II treats communication through a limit price signalling game. The examples themselves are quite stylized and are not intended to be realistic. Neither the technical aspects of the models nor the often considerable modelling efforts that went into them are discussed in any detail. Instead, the examples portray a conceptual framework in which to view this new body of work.

I. Binding Commitments and Self-Enforcing Threats

This section examines a class of models[1] in which information is perfect and communication costless. The models themselves differ in

*Federal Trade Commission and University of Pennsylvania. I would like to thank Rich Gilbert, Bob Reynolds, and my globetrotting collaborator Joe Stiglitz for helpful conversations, and Penny Campbell for excellent typing and editing. The remarks in this statement represent only my personal views. They are not intended to be, and should not be construed as, representative of the views of any other member of the Federal Trade Commission staff or individual Commissioners.

[1]See Donald Hay; B. Curtis Eaton and Richard Lipsey; A. Michael Spence (1977); Edward Prescott and Michael Visscher; Richard Schmalensee; and others cited in the text. Some rather subtle first-mover advantages are discussed in Oliver Williamson (1975).

the deterrence instruments (capitalization rate, brand selection, innovation and advertising, for example), the preentry market structure (monopoly, cartel, Nash-Cournot oligopoly) and the rules governing postentry interaction (Nash-Cournot, Stackleberg leadership "predatory" $P = MC$). They also differ in central focus from existence of a positive profit free-entry equilibrium to alternative antipredation policies. Their similarity is the use of binding commitments.

Absent entry, suppose an established firm (the monopolist) can earn a positive (excess) present discounted value of profits v_0. Given some assumed (and agreed upon) rule governing postentry interaction (the equilibrium concept), the monopolist and equally efficient large-scale entrant would each earn a lower $v_1 < v_0$. If $v_1 > 0$, entry will clearly occur.

Alternatively, suppose the monopolist may select a preentry expenditure level c that the entrant must exactly match in order to survive. This expenditure may be the selection of any of the instruments listed previously. For expository purposes, if the expenditure is viewed as an excess of brand-awareness advertising, then the equal matching requirement represents an assumption of equal advertising efficiency. Will the monopolist advertise, and if so, at what level? The outcome matrix is shown in Table 1, with the entrant's returns tabulated first.

By advertising at a level $c \geq v_1$ the entrant is deterred by the prospect of nonpositive profits, and the monopolist earns $v_0 - c$. Without advertising, entry occurs and the monopolist earns v_1. Thus, at the minimum deterrence level $c = v_1$, deterrence is profitable if $v_0 \geqq 2v_1$. As Richard Gilbert notes, if industry joint profits are maximized with only a single existing firm,[2] then this profitability condition is met.

This example illustrates the essential feature of this approach to strategic entry deterrence. The necessity of a binding commitment is obvious; if on-going control of the advertising were possible, an entrant would rationally forecast that the monopolist

TABLE 1—MATRIX OF OUTCOMES

Entrant	Monopolist No Ads	Monopolist c Ads
Entry (matching ads)	v_1, v_1	$v_1 - c, v_1 - c$
No Entry	$0, v_0$	$0, v_0 - c$

will curtail the mutually destructive ads and accomodate him once he enters.

Gilbert, Gilbert and David Newberry, Robert Reynolds, and Joseph Stiglitz, Gilbert, and Partha Dasgupta apply a variant of the argument to preemptive innovation. An established firm earning v_0 may deter entry by innovating first. If the entrant innovates first, the monopolist and entrant each earn $v_1 < v_0$. Thus, the monopolist's net benefit $B_m = v_0 - v_1$ always exceeds the entrant's $B_e = v_1$ as long as the maximum joint-profit condition mentioned previously is met.

All these stated postentry outcomes represent reduced forms to hypothetical postentry games. As Williamson (1977) and William Baumol point out, the rules of the game, in the form of antitrust limitations placed on postentry responses, are crucial. Rules more favorable to entry may be at the cost of limiting innocent competitive adjustments The rules themselves may be an area of strategic planning. In Spence (1977), it is in a cartel's interest to ensure its self-destruction in the face of entry; that policy decreases the gains to entry and, hence, improves deterrence.

Elsewhere I have investigated another variant that pays explicit attention to the postentry game—the self-enforcing threat game. By altering his postentry incentive structure, a monopolist may manipulate the outcome of the game. For example, suppose the monopolist has two postentry output possibilities q_1 and q_2 that lead to payoffs of $v_1 > 0$ and $v_2 < 0$ for both monopolist and equally efficient entrant. Hence, the entrant correctly forecasts q_1 and successfully enters.

However, suppose the monopolist permanently adopts a different *inferior* technology with everywhere lower payoffs $v_i' < v_i$ to himself, though not to the entrant. By judiciously choosing a cost structure that reverses

[2]The condition is weak since the monopolist has the choice to establish his own noncompeting subsidiary.

the relative payoffs,[3] that is $v_2' > v_1'$, a postentry incentive is created to choose the mutual loss-producing output q_2. Since q_2 entails negative profits $v_2 < 0$ to the entrant, he is deterred. This self-enforcing threat strategy is profitable if the alternative technology is not too inferior, that is, if its maximum no-entry payoff exceeds the maximum postentry payoff v_1 under the superior technology.

II. Limit Pricing as a Signal

Established firms and potential entrants do not have perfect information about the outcome matrix. Deterrence and entry represent risky investments. The entrant must rely on observable variables as signals for the actual relevant data. Joe Bain considers the use of current (limit) price as a signal for postentry price intentions. In this spirit, Robert Rosenthal and Reinhardt Selten explore the incentive of a multimarket or multiperiod monopolist to maintain low prices in one market or period to create a deterrence reputation in others. Michael Porter discusses a number of elements of the signalling process.

In this context, established firms have an incentive to manipulate the signalling process. James Friedman and Reynolds and I have studied examples in which the current price signals the established firm's cost structure to the entrant; that is, the entrant indirectly infers marginal cost from price observations. Complexity is introduced because lower-cost established firms wish to maintain the veracity of the signal, whereas higher-cost firms desire confusion. The outline of an example with self-fulfilled expectations is as follows.

Suppose that across a set of industries with identical demand elasticities, some are monopolized by low-cost and the rest by high-cost monopolists. A potential entrant with intermediate cost may attempt entry into a single industry. If the monopolists engage in innocent profit maximization, then the entrant is able to perfectly infer cost differentials from the price differentials, since all demand elasticities are identical.

It is in the interest of the high-cost producers to charge the low price to fool the entrants. However, at identical prices, the low-cost producers will charge a lower price to reestablish the signal; even these low-cost producers must bear short-run costs to drive out entrants. According to the proportion of high- and low-cost monopolies, the number of potential entrants, the cost of driving out unsuccessful entrants and the cost differentials themselves, either a single price "pooling" equilibrium or a dual price "separating" equilibrium will obtain.

In this very simple example, in a pooling equilibrium no entry occurs; for if it did, a low-cost monopolist could easily deter entry from his market by altering his price slightly to reestablish the signal. Similarly, a separating equilibrium entails entry, for the signal is maintained. It is also possible, following Bain, Rosenthal, and Selten, that a mixed strategy of randomizing prices (an entry-deterring noisy monopolist) will obtain. In a more complicated model with many types of firms with different demand elasticities and costs, some pooling with entry should occur. Reasoning from other signalling models, (see Spence, 1973; Michael Rothschild and Stiglitz) there will be overinvestment in the signal; that is, limited pricing strategies will be adopted.

III. Conclusions

These various examples have the common feature that the monopolist creates an incentive to choose postentry actions to the detriment of entrants. Technically, this incentive is created by irreversibly altering his profit function. In a personal communication Stiglitz pointed out that in general all deterrence instruments create intertemporal relationships in the profit function. In this sense, all deterrence instruments act as "capital," and a binding commitment corresponds to "irreversible investment." In limit-pricing models, the preentry price is effectively converted to capital by its role in forming the basis of the entrant's expectation of costs, elasticity, and conjectural variations.

[3]Such a reversal is not unreasonable Suppose $q_2 > q_1$ and the alternative technology has increased fixed costs and decreased marginal costs.

REFERENCES

J. Bain, "A Note on Pricing in Monopoly and Oligopoly," *Amer. Econ. Rev.*, Mar. 1949, *39*, 448–64.

W. J. Baumol, "Quasi-Permanence of Price Reductions: A Policy for Prevention of 'Predatory Pricing'," unpublished paper, New York Univ. 1978.

B. C. Eaton and R. G. Lipsey, "The Theory of Spatial Pre-Emption: Location as a Barrier to Entry," unpublished paper, Queens Univ., Feb. 1976.

J. W. Friedman, "Limit Price Entry Prevention when Complete Information is Lacking," unpublished paper, Univ. Rochester 1978.

R. J. Gilbert, "A Note on Preemptive Competition," unpublished paper, Univ. California-Berkeley 1978.

——— **and D. Newberry,** "Preemptive Innovation," unpublished paper, Univ. California-Berkeley 1978.

D. A. Hay, "Sequential Entry and Entry-Deterring Strategies in Spatial Competition," *Oxford Econ. Papers*, July 1976, *28*, 240–57.

M. E. Porter, "Market Signals," unpublished paper, Harvard Bus. Sch., 1978.

E. C. Prescott and M. Visscher, "Sequential Location Among Firms with Foresight," *Bell J. Econ.*, Autumn 1977, *8*, 378–93.

R. J. Reynolds, "Entry Reaction and Preemptive Product Innovation," unpublished paper, U.S. Department of Justice 1978.

——— **and S. C. Salop,** "Credible Limit Pricing and Entry," unpublished paper, U.S Department of Justice 1978.

R. W. Rosenthal, "Games of Perfect Information, Predatory Pricing and The Chain-Store Paradox," unpublished paper, Bell Laboratories 1978.

M. Rothschild and J. Stiglitz, "Equilibrium in Competitive Insurance Markets: The Economics of Imperfect Information," *Quart J. Econ.*, Nov. 1976, *90*, 629–49.

S. C. Salop, "A Note on Self-Enforcing Threats and Entry Deterrence," unpublished paper, Univ. Pennsylvania 1978.

Thomas C. Schelling, *The Strategy of Conflict*, Cambridge, Mass. 1960.

R. Schmalensee, "Entry Deterrence in the RTE Cereal Industry," *Bell J. Econ.*, forthcoming.

Reinhardt Selten, "The Chain-Store Paradox," unpublished paper, Universität Bielefeld 1974.

A. Michael Spence, *Market Signalling*, Cambridge 1973.

———, "Entry, Capacity, Investment and Oligopolistic Pricing," *Bell J. Econ*, Autumn 1977, *8*, 534–44.

J. Stiglitz, R. Gilbert, and P. DasGupta, "Invention and Innovation Under Alternative Market Structures: The Case of Natural Resources," unpublished paper, Oxford Univ. 1978.

Oliver E. Williamson, "Predatory Pricing: A Strategic and Welfare Analysis," *Yale Law J.*, Dec. 1977, *87*, 284–340.

———, *Markets and Hierarchies*, New York 1975.

Equilibrium in Product Markets with Imperfect Information

By J. E. STIGLITZ*

This paper is concerned with the relationship between information and market equilibrium: with the effect of information on the effective degree of competition, on the level of prices and their dispersion, on the variety and character of products produced by markets, on the one hand; and with the demand for information by consumers and the supply of information by producers, on the other. I shall argue not only that taking appropriate account of the costs of information provides an explanation of many phenomena which otherwise could not be explained, but that it also casts considerable doubt on a number of presumptions of traditional economics, for example, the efficiency of competition, and the policy prescriptions derived from those presumptions.

Traditional models of competition with perfect information obviously cannot explain the widely observed phenomena of price distributions, which seem sufficiently persistent that they cannot simply be dismissed as a disequilibrium phenomenon; nor can they explain advertising; nor can they explain why markets in which there are only a few large firms often seem more competitive than markets with many small firms. The work I am about to describe, which attempts to characterize equilibrium in product markets in which information is costly, provides considerable insight into these phenomena. At the same time, examining markets with costly information raises several important conundrums for competitive equilibrium theory: in the simplest of models formulated, no market equilibrium exists. The resolution of this paradox provides one of the foci of this paper.

It is important to recognize that market structure is itself an endogenous variable, a result (at least in many cases) of natural barriers to entry and incentives to agglomerate, some of which are related in an essential way to the cost of information. In this paper, I shall not have time to explore this set of relationships; throughout, it shall be assumed that the only barrier to entry is the fixed cost of establishing a new firm and that these are sufficiently small that markets will be characterized by a large number of firms.

I. Imperfect Information and Monopoly Power

A. *Competitive Markets with Monopoly Price*

It has long been recognized that imperfect information would result in firms having some degree of monopoly power. But it was presumed that a market equilibrium with "small" costs of search would be "very similar" to one with zero search costs; I suspect this belief is based on the presumption of continuity which seems to have been ingrained in economists at least since Alfred Marshall's famous dictum.

As Tibor Scitovsky and Peter Diamond have shown, this is not necessary; even with infinitesimal search costs and a very large number of firms, in market equilibrium *the price is the monopoly price.* To see this, assume all individuals have the same demand curve, all firms have the same cost functions, and also all charge the same price. Then if price is below the monopoly price, it would pay any firm to raise its price by an amount less than the magnitude of search costs for the individual with the lowest search cost, for it would then not pay any customer to leave the given store and go to another. But this, in turn, implies that all firms will raise their

*Professor of economics, All Souls College, Oxford University. This paper was written while I held the Oskar Morgenstern Distinguished Fellowship at Mathematica and was visiting professor at the Institute for Advanced Study. Research support from the National Science Foundation is gratefully acknowledged. I am deeply indebted to Steven Salop for helpful discussions concerning the topics of this paper; for a parallel nontechnical survey of these topics, see Salop (1978a).

prices until marginal revenue equals marginal costs.

So far, nothing has been said about the determination of the number of stores. If we allow free, competitive entry, we obtain a result of considerable importance: *all the monopoly profits are dissipated in excess entry; competition is socially wasteful. Monopoly is Pareto superior to free competition.*[1]

This is not the only presumption of conventional economic theory which needs to be reversed. In market equilibrium, if there are a large number of firms, no single firm has any effect on the search behavior of individuals. But if a group of firms can get together, they can, by lowering their price, induce individuals to search to find one of the members of the "low-price chain." Thus a reduction in the number of competing firms may result in more *effective* competition and lower prices to the consumer.

B. *Nonexistence of Equilibrium: Gresham's Law Revived?*

Among businessmen, there is some belief in the converse proposition, that excessive competition may actually destroy markets in which there is costly information. This can occur even in the simplest of search models. For instance, assume that individuals always purchase one unit of the good (if they purchase it), and that the dollar value of the utility they derive from it is equal to u. Hence, when consumers arrive at a store, they purchase if and only if the price p is less than u. In this context, the monopoly price is u, and hence each firm will raise its price p to u. But if it is costly to go to the market, then the utility from entering the market itself is $u - p - c = -c < 0$, where c is the cost of search. Thus, no consumer will enter the market. In their ruthless attempts to exploit the hapless consumer who has entered the market, firms continue to raise the price to the point where the marginal individual who has entered the market is just indifferent to purchasing or not purchasing; but if all of his consumer's surplus is eliminated, then he will have no incentive to enter the market at all. But since each firm is small, it perceives itself to have no effect on the individual's decision to enter the market.[2] Hence, if all individuals have *positive search costs* (no matter how small), *there never exists a competitive market equilibrium with uniform prices for homogeneous commodities*. This result seems a quite general and disturbing paradox. Three possible resolutions are addressed in Sections II, III, and IV.

II. Equilibrium Price Distributions

Market equilibrium may be characterized by a price distribution. Although the existence of price distributions provides one of the main motivations for market search (see George Stigler), it initially apeared difficult to formulate a consistent model in which price dispersion persisted. To do so, as I pointed out in my forthcoming paper, it is necessary to explain (a) why individuals do not eventually learn about the prices charged in any particular store; and (b) why different stores charge different prices. There are some simple, if not completely convincing, explanations for both of these phenomena. A flow of ignorance can be maintained either by entry of new firms or new individuals. If firms differ with respect to their cost functions, then the price they set will differ, and the market will give rise to a price distribution (see Salop, 1973). Alternatively, if spatially separated markets are subjected to random shocks and are imperfectly arbitraged, there will also be a price distribution (see, for example, Dale Mortensen).

More interesting, however, are situations where firms have identical cost functions. Then it must be the case that the profits

[1]It is important to emphasize that these considerations have to be balanced off against other, probably more important, advantages of competition, which are not well captured by the traditional analyses of competitive markets. See the author (1978a).

[2]The result (see the author, 1977a) is general: individuals may have conventional downward-sloping demand curves which may differ from individual to individual, search costs may differ from individual to individual, etc. In these situations, the firm will employ non-linear price schedules. (See the author 1977b.)

generated at one price must be the same as those at another price. Not only must profits as a function of price have several peaks but those peaks must be of exactly the same height. This seems remarkable, until it is recognized that the shape of the profit function is itself a function of the price distribution.

In particular, high-price stores may have higher average costs, either because they spend more to acquire customers or because they have fewer sales (with U-shaped average cost curves); the differences in costs exactly offsetting the difference in price. Several such models have been constructed.

A. *Bargains and Ripoffs: Markets Exploit High-Search-Cost Individuals*

Salop and I (1977) have investigated a model in which the low-price stores have higher sales because individuals with low-search cost seek them out. Only high-search-cost individuals go to high-price stores. This model with costly information establishes the existence of market equilibrium with price dispersion, and illustrates several further general principles.

First, *the market itself gives rise to the imperfection of information.* In a "planned" economy, all firms would charge the same price, and hence there would be perfect information. Second, *the widespread belief that all that is required for markets to work well is that there be some individuals who are informed and who thus arbitrage the market is not valid,* at least within this context. The informed convey only a limited positive externality on the uninformed, in the sense that without them there would be fewer stores charging the competitive price. However if there are enough informed individuals, the market may act competitively. On the other hand, an increase in the number of uninformed creates a negative externality: there will now be a larger proportion of high-price stores, and the expected welfare of all the uninformed is decreased.

Indeed, the *market equilibrium in which all individuals have costly information must be characterized either by firms charging the monopoly price or by price dispersion.* For it pays individuals to undertake information acquisition only if there is price dispersion, that is, the market is not fully arbitraged: so if there is no price dispersion, all individuals remain uninformed, and the firms can then exercise their monopoly power.

B. *The Theory of Sales*

In the model just described differences in individuals' information costs give rise to the price distribution. Salop and I also considered a "Theory of Sales," in which all individuals and all firms are identical and still, the only equilibrium may be a price distribution (1976). This model recognizes that firms can increase sales both by attracting the more informed, and by selling more to those who arrive, for example, by inducing them, "in a sale," to purchase for future consumption (storage). If all stores charged the same price, it would pay some firm to lower its price to induce those who are planning to consume the good next period as well to purchase for their future needs. They lower their price just to the point necessary to induce purchase for storage. Equilibrium is established where the extra sales just compensate for the lower price on each sale.

It is important to observe that in these cases, where firm profits from charging a high price or a low price are the same, there is an alternative interpretation of the equilibrium:[3] each store randomizes its price between the high price and the low price. In that case, even if there were no entrants into the market, or new firms, individuals would remain imperfectly informed.[4]

Although it is now apparent that the existence of markets may, even in equilibrium, be characterized by price dispersion, this does not fully resolve the paradox of nonexistence for two reasons. First, in the sequential search model (where at each search individuals only

[3]Since the profits at the two prices are the same, the firm is indifferent which price it charges. Equilibrium is thus characterized by a mixed strategy. This interpretation is only persuasive if the *ex post* cost function is the same as the *ex ante* cost function.

[4]For a similar model with mixed strategy equilibrium, see Y. Shilony.

find about the store they have visited) if all firms are identical, then such markets can never be characterized by price dispersion. For clearly, if there were a price distribution, the reservation price (the price above which the individual does not buy) must be slightly above the lowest price within the distribution. It would thus pay that store to raise its price, provided that the price were below the monopoly price. Secondly, even if firms and individuals differ, by the same kind of reasoning as employed earlier, it will always pay all stores to raise their prices to the point where the marginal entrant into the market (the individual who is just indifferent to purchasing or not) is indifferent to purchasing (given that he has already entered). But then, he will not want to enter the market. Hence, there cannot exist a "marginal" entrant, and hence there cannot exist a market.

III. Product Heterogeneity

A. *Quality Dispersion*

The second possible resolution to the nonexistence paradox is that markets are characterized by product heterogeneity. Of course some product heterogeneity is really nothing more than disguised price dispersion. A good x which lasts twice as long as a good y (if the interest rate is zero) is just equal to two units of y. There is, however, one important difference between price and quality dispersion: even after arriving at the store, the individual may not know what the true price of the commodity is. Only after purchasing the commodity does he find out the effective (true) price.

Obviously, this lack of knowledge provides an incentive for firms to "cheat," to lower their quality, thereby raising their effective price. This is limited by two considerations: firms can establish a reputation if the commodity is repeatedly purchased. Then "good" firms will have "repeat customers" and thus larger sales. Secondly, firms can provide guarantees, which both reduce the risk borne by the customer (see Geoffrey Heal) and serve as a "signal" (a self-selection device) by the firm concerning its quality.

Even if firms are risk neutral, the use of guarantees is limited by both moral hazard and adverse selection problems—a 100 percent guarantee will result in the purchaser abusing the product (if he can), and if use cannot be monitored, will lead to purchase by those individuals who are "hardest" on the commodity. Moreover, guarantees have to be guaranteed, and the enforcement of such contracts may be costly.

Individuals can attempt to infer something about the product (or the seller) not only on the basis of the guarantee provided but also on the basis of the price charged. George Akerlof, in his classic theory of lemons considered a particular variant of this problem. He assumed that the average quality z of a commodity offered on the market was an increasing function of the price $z(p)$, while demand was a function of price and average quality. Since quality decreases with price, it is possible that the "effective price," the price per unit quality p/z, actually increases as the nominal price decreases. In Akerlof's model, the only possible equilibrium in which demand equaled supply was where price was zero: quality was zero, so demand was zero, and supply was zero.

Subsequently, the author (1976), Salop (1979), the author and Andrew Weiss, and Charles Wilson have argued that, at least in some contexts, this is an inappropriate equilibrium concept. When individuals know that quality is affected by price either because of screening or incentive effects, they will use price to affect the expected quality of the commodity purchased. For instance, in the labor market, a firm will choose the wage to minimize expected labor costs, that is, minimize the wage per effective labor unit; in the capital market, a lender will choose the interest rate to maximize the expected return obtained from the loan; and in the product market, with which we are concerned here, the buyer will choose p to minimize the "quality adjusted price."

In this case, competitive market equilibrium does not necessarily entail market clearing. The law of supply and demand has been

repealed. Assume that at the value of p at which the quality adjusted price p/z is minimized, there is some individual willing to supply a good. Conventional theory has it that he undercuts his rivals; price is thus bid down until the market is cleared. But here, if an individual offers the commodity for sale at a lower price, then the buyer infers that it is probably of a lower quality, sufficiently lower in fact not to compensate for the lower price. The existence of non-market-clearing equilibria has important implications for macroeconomic analysis (see Salop, 1979; the author, 1978b).

B. *Product Variety*

In the discussion so far, all individuals agree on the desirable characteristics of a commodity. But in many situations, different individuals place different relative evaluations on the different goods offered in the market, that is, there is some element of matching individuals to commodities (or jobs). Assume, for instance two kinds of widgets, blue widgets and red widgets. Each store can only sell one kind. Some individuals prefer blue widgets, some red. Store owners cannot discriminate between blue widget lovers and red widget lovers. Then, market equilibrium may be characterized by a price dispersion, with low-price stores selling to those who like their commodity *and* those who would prefer the other color, but, given the low price, are willing to stop searching, while the high-price stores sell only to those who are "properly matched." More importantly for our purposes here, there may exist a market equilibrium: an individual who goes to a low-price store which sells the commodity which he loves enjoys some consumer's surplus: it is the possibility of capturing this consumer's surplus which induces individuals to bear the costs of entering the market.

Although this leads to the result that market equilibrium will be characterized by some product variety, there is no presumption that the market will have the correct amount of variety. Even with perfect information, of course, markets will not, in general, have the correct amount of product differentiation. (See Michael Spence, Avinash Dixit, and the author, 1977a.) But the existence of costly information changes both the social and private returns to variety. Those who sample the new commodity and don't like it are worse off, since they now have to search more than they otherwise would. The potential beneficiaries of the new commodity are those for whom there is now a commodity more to their liking, but because of costly search, only a fraction of these individuals actually get the new commodity.

At the same time, the existence of product variety may have an important effect on the returns to improved information. For improved information may affect the elasticity of demand facing any firm; if it increases the degree of monopoly power of each firm, it may lead to higher prices and lower welfare. If it lowers the degree of monopoly power, not only will there be better matching of individuals and goods, but markets may be more competitive and prices lower.

C. *Kinked Demand Curves Arising from Imperfect Information*

The models formulated so far are basically static, but one important implication of the natural dynamic extension is that demand curves are kinked. To see this, assume a market in equilibrium, with the only individuals searching for the lowest price (or the commodity most to their liking) being new entrants. Then a firm which raised its price would induce the marginal individuals purchasing it to begin searching again, and it thus loses customers; but when it lowers its price, since individuals at other stores do not know about its lower price, it does not gain a corresponding number of new customers. It should be noted that this argument for the kinked demand curve is distinctly different from the traditional oligopoly argument, which postulates (not completely convincingly) asymmetric responses on the part of rivals (assumed to be few in number—here there may be a large number of firms) to price increases and decreases. The fact that

demand curves (and by a similar argument, labor supply curves) are kinked has important implications for macro-economic equilibrium (see Salop and the author, 1976; the author, 1978b).

IV. Advertising

The third resolution of the nonexistence paradox is that individuals can obtain information by means other than sequential search.[5] Earlier, we described one such method, where individuals purchased a newspaper, obtaining complete information. Some individuals hear about "good stores" by word of mouth and this too may lead to a price distribution (see Salop and the author, 1976). In both cases, the existence of some low-price stores with high sales provides an incentive for some individuals to enter the market. But firms can attempt to provide information to their customers as well, through advertising.

This, however, gives rise to a new advertising paradox. Assume a set of firms advertise a price p. Then, it always pays any firm to advertise at a slightly lower price, for it will obtain all the customers. This means that advertised prices continue to be cut, until they reach the minimum average cost. But then, firms must be making a loss, since they are breaking even on their sales, but paying advertising costs.

Gerard Butters provided an ingenious resolution of the advertising paradox: he showed that there existed a continuous equilibrium price distribution. Stores randomly send out advertisements informing customers of their price. The individual goes to the cheapest store from which he receives an advertisement. Thus, most of the advertisements of high-price stores are ineffective; while all the advertisements of the lowest-price store are effective. The higher sales of the lower-price store exactly compensates for the lower sales of the higher-priced stores.

Although there are some markets in which high-priced commodities have a higher expenditure on advertising, in other markets (discount department stores), the low-priced stores have larger advertising budgets.

There is an alternative resolution of the "advertising paradox." Assume, as I did earlier, that markets are characterized by product and consumer heterogeneity. Then a store can actually advertise and raise its price, since individuals who like the given commodity would rather buy there than attempt the random lottery of the market. Equilibrium of such markets will in general have price dispersion; low-price stores may advertise a great deal, but will make up for both the lower price and the higher advertising costs with the larger sales arising from the wider market area they serve.

V. Concluding Remarks

Although space prohibits pursuing all the implications of imperfect information for equilibrium in product markets, I hope I have convincingly shown that the traditional paradigms of competitive markets, with perfect information and markets equilibrated by the mythical Walrasian auctioneer, are not only not directly applicable, but may be seriously misleading. For instance, attempts to promote competition by increasing the number of firms by removing barriers to entry may actually reduce effective competition, increase prices, and lead to lower efficiency. Similarly, although clearly the traditional presumption of caveat emptor has no basis within welfare economics when information is costly, the full implications of various attempts at consumer protection need to be examined carefully within a well-articulated model of product market equilibrium of the kind I have attempted to formulate here before their desirability can be correctly assessed.

Finally, the kinds of models I have developed here do seem to provide, at last, a micro-economic foundation for many important macro-economic phenomena: not only have I shown how they can give rise to rigidities in adjustment but I have also shown

[5]Equivalently, it may be assumed that there is a mass of individuals with zero search costs. This ensures that at least some stores charge the competitive price; if there is a distribution of individuals by search costs, a price distribution may then be generated. See Peter von zur Muehlen.

how there may be equilibria in which markets do not clear.

REFERENCES

G. A. **Akerlof**, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quart. J. Econ.*, Aug. 1970, *84*, 488–500.

G. **Butters**, "Equilibrium Distributions of Sales and Advertising Prices," *Rev. Econ. Stud.*, Oct. 1977, *44*, 465–92.

P. A. **Diamond**, "A Model of Price Adjustment," *J. Econ. Theory*, June 1971, *3*, 156–68.

A. **Dixit and J. E. Stiglitz**, "Monopolistic Competition and Optimal Product Differentiation," *Amer. Econ. Rev.*, June 1977, *67*, 297–308.

G. **Heal**, "Guarantees and Risk Sharing," *Rev. Econ. Stud.*, Oct. 1977, *44*, 549–60.

D. T. **Mortensen**, "Search Equilibrium in a Simple Multi-Market Economy," disc. paper no 54, Center Math. Stud., Northwestern Univ., Oct. 1973.

S. **Salop**, "Wage Differentials in a Dynamic Theory of the Firm," *J. Econ. Theory*, Aug 1973, *6*, 321–44.

——, (1978a) "Parables of Information Transmission," in A. Mitchell, ed., *The Effect of Information on Consumer and Market Behavior*, Chicago 1978.

——, (1978b) "Second Best Policies in Imperfect Competition," unpublished paper, Univ. Pennsylvania 1978.

——, "A Model of the Natural Rate of Unemployment," *Amer. Econ. Rev.*, Mar. 1979, *69*, 117–25.

—— **and J. E. Stiglitz**, "Search Costs, Monopoly Power, and Price Distributions," mimeo., Stanford Univ. 1976.

—— **and** ——, "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion," *Rev. Econ. Stud.*, Oct. 1977, *44*, 493–510.

T. **Scitovsky**, "Ignorance as a Source of Oligopoly Power," *Amer. Econ. Rev.*, Mar. 1950, *40*, 48–53.

Y. **Shilony**, "Mixed Pricing in Locational Oligopoly," *J. Econ. Theory*, Apr. 1976, *14*, 373–88.

M. **Spence**, "Product Selection, Fixed Costs, and Monopolistic Competition," *Rev. Econ. Stud.*, June 1976, *43*, 217–35.

G. **Stigler**, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–25.

J. E. **Stiglitz**, "Equilibrium Wage Distribution," *Econ. J.*, forthcoming.

——, "Prices and Queues as Screening Devices in Competitive Markets," IMSSS tech. rept. no. 212, Stanford Univ., Aug. 1976.

——, (1977a) "Some Rough Notes on Diversity of Tastes and Diversity of Commodities," paper presented at Bell Laboratories Conference on Information and Market Structure, Feb. 1977.

——, (1977b) "Monopoly, Non-linear Pricing and Imperfect Information: The Insurance Market," *Rev. Econ. Stud.*, Oct. 1977, *44*, 407–30.

——, (1978a) "Information and Competition," Inaugural Lecture presented at All Souls College, 1978.

——, (1978b) "Lectures in Macro-Economic Theory," lecture 6, mimeo., Oxford Univ. 1978.

—— **and A. Weiss**, "The Theory of Credit Rationing and Imperfect Information," paper presented to Honolulu Meeting of Western Economic Association, June 1978.

P. **von zur Muehlen**, "Limited Price Information and Monopolistic Competition," mimeo., Fed. Reserve Board 1976.

C. A. **Wilson**, "Equilibrium and Adverse Selection," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 313–17.

Multiproduct Technology and Market Structure

By ROBERT D. WILLIG*

A recent line of research has exposed some technological determinants of the structure of industries that produce more than one good. The analyses of both multiproduct perfect competition and natural monopoly require a generalized notion of average cost and, in addition, several newly identified technological characteristics pertinent only to joint production. This paper provides an overview of these new results, and suggests a unifying framework in which the theory can be further developed.

I. Economies of Scope[1]

There are two basic reasons to study multiproduct firms. First, casual empiricism suggests that there are virtually no single product firms. Second, the technological characteristic we have named economies of scope may force firms in industry equilibrium to produce more than one good.

There are economies of scope over the production of goods 1 and 2 if[2]

$$(1)\quad C(q_1, q_2) < C(q_1, 0) + C(0, q_2), \qquad q_1 > 0, q_2 > 0$$

where $C(q_1, q_2)$ is the firm's (minimized) cost of producing q_1 units of good 1 jointly with q_2 units of good 2, at given parametric input prices (which the present notation suppresses). With economies of scope, joint production of two goods by one enterprise is less costly than the combined costs of production of two specialty firms. The degree of economies of scope is

$$(2)\quad Sc \equiv [C(q_1, 0) + C(0, q_2) - C(q_1, q_2)]/C(q_1, q_2)$$

This is the proportion by which costs would increase if a multiproduct firm were split along product lines. Rewriting (1) as $C(q_1, q_2) - C(q_1, 0) < C(0, q_2)$ shows that with economies of scope, the cost of adding the production of q_2 to the production of q_1 is smaller than the cost of producing q_2 alone The smaller incremental cost of product line 2 is enjoyed by the enterprise with the broader scope of production.

This effect may cause economies of scope entry barriers. Entry into oligopolistic industries may be more difficult and more risky the greater the number of new products that the entrant must develop, market, and tool up to manufacture (see Richard Schmalensee) Yet, the greater the degree of scope economies, the larger the cost disadvantage of a firm that offers fewer products. Moreover, the wider the set over which economies of scope extend, the more product lines a potential entrant must plan in order to avoid these cost disadvantages.

Economies of scope arise from inputs that are shared, or utilized jointly without complete congestion. The shared factor may be imperfectly divisible, so that manufacture of a subset of the goods leaves excess capacity in some stage of production. Or some human or physical capital may be a public input which, when purchased for use in one production process, is then freely available to another. This limiting case is useful to model. Let $V^i(q_i, K)$ be the cost of producing q_i when K units of the public input are available. When the price of a unit of the public input is r, the multiproduct cost function is

$$(3)\quad C(q_1, q_2) = \min_K [V^1(q_1, K) + V^2(q_2, K) + rK]$$

*Professor of economics and public affairs, Princeton University. The title of this paper will be shared with a book by W. Baumol, J. Panzar, and myself.

[1]This term was coined in John Panzar and the author (1975). This section follows Panzar and the author (1978). See William Baumol and Yale Braunstein for an empirical study.

[2]The cited papers cover the general case of n products.

This yields cost functions with economies of scope. For example, if $V^i(q_i, K) \equiv f^i(q_i)g(K)$, $g < 0$, then

$$(4)\quad C(q_1, q_2) = rT\left[\frac{f^1(q_1) + f^2(q_2)}{r}\right], \quad T' > 0, T'' < 0$$

II. Multiproduct Returns to Scale and Ray Average Cost[3]

A natural generalization of single product average cost is the ray average cost of a multiproduct firm:

$$(5)\quad RAC(q) = \frac{C(q)}{\Sigma q_i} = \frac{C(tq^0)}{t}$$

where $tq^0 = q$ and $\Sigma q_i^0 = 1$. This is the standard average cost of the composite commodity whose unit is the vector q^0. In the construction of a ray average cost curve, the output point moves along a ray through the origin of output space; hence the name. For the usual reasons, one would suppose that ray average cost curves have a U shape.

The degree of scale economies at q, denoted by S, is equal to $1/(e+1)$, where e is the elasticity of the relevant ray average cost curve at q. The variable S is greater than, less than, or equal to one as returns to scale are locally increasing, decreasing, or constant, and as the ray average cost curve's derivative is negative, positive, or zero. If the technology can be represented by a differentiable transformation function $\phi(x,q)$ which is nonnegative if outputs q can be produced from inputs x, then $S = -[\Sigma x_i \partial\phi/\partial x_i]/[\Sigma q_j \partial\phi/\partial q_j]$, where the derivatives of ϕ are evaluated at q and at the cost efficient value of x. Also, in this case, if all inputs are expanded by some small factor λ then all the outputs can essentially be expanded by the factor λ^S. Thus S can be interpreted as the *local* (in that it can vary with q) degree of homogeneity of the technology. Finally, $S = C(q)/\Sigma q_i C_i(q)$, the ratio between production cost and the revenues that would accrue from pricing the outputs at their marginal costs. Thus, these revenues exceed, fall short of, or equal cost as there are decreasing, increasing, or locally constant returns to scale.

III. Product-Specific Returns to Scale[4]

The incremental cost of a product line is the extra cost of adding the production of that line to the other outputs of the firm. The average incremental cost curve for, say, product 1 plots $AIC_1(q) \equiv [C(q_1, q_2) - C(0, q_2)]/q_1$ against q_1 for fixed q_2. Any setup costs specific to product 1 yield decreasing average incremental costs ($DAIC_1$) for small q_1. If plant used to produce q_2 becomes increasingly congested as q_1 rises, then there may be increasing average incremental costs ($IAIC_1$) beyond some level of q_1. The degree of product-specific returns to scale at q is defined as $S_i = 1/(1+e_i)$, where e_i is the elasticity of $AIC_i(q)$ with respect to q_i. Also $S_i = AIC_i(q)/C_i(q)$. Thus, the average incremental and marginal cost curves are interrelated in the same way as are single product average cost and marginal cost curves. There is said to be increasing (decreasing) returns to scale in product i if $S_i > 1$ ($S_i < 1$) which in turn implies $DAIC_i$ ($IAIC_i$).

The degrees of economies of scope, and overall and product-specific returns to scale are inextricably related:

$$(6)\quad S = \frac{w_1 S_1 + w_2 S_2}{(1 - Sc)}$$

where $w_i = q_i C_i(q)/\Sigma q_j C_j(q)$; S is a weighted average of S_1 and S_2, magnified by $1/(1 - Sc)$. Thus, economies of scope and $DAIC$ in each product line together imply overall scale economies.

Figure 1 displays, for a particular technology with economies of scope, the sets of output vectors at which overall returns, and returns to each product line are increasing

[3]This treatment of returns to scale follows Panzar and the author (1977b). The ray average cost concept was developed in Baumol (1976, 1977).

[4]This term was coined by F. M. Scherer et al. The definition and development here largely follow that in Panzar and the author (1978, 1977a).

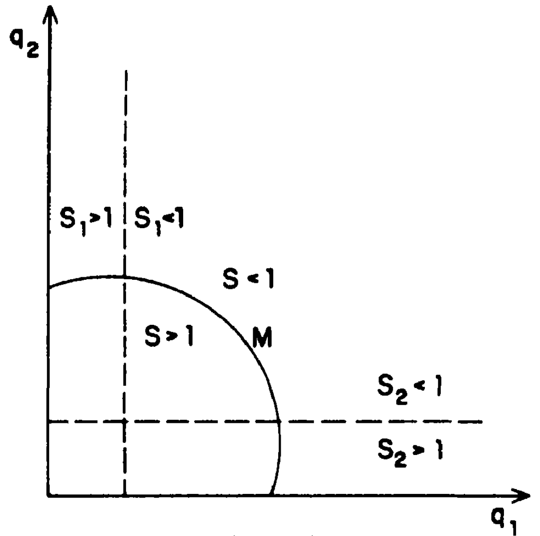


FIGURE 1

(decreasing). Note that S can exceed 1 with $S_1 < 1$ and $S_2 < 1$ since $Sc > 0$. As drawn, S_1 is independent of q^2 and S_2 is independent of q_1. This occurs iff $C(q_1, q_2) = C(q_1, 0) + C(0, q_2) - aC(q_1, 0)\, C(0, q_2)$.

IV. Transray Convexity of Costs[5]

A cost function is transray convex along a hyperplane $\Sigma a_i q_i = k$, $a_i > 0$, if for any vectors of outputs q_*, q_{**} on that hyperplane and for $0 \leq \lambda \leq 1$,

$$(7)\quad C(\lambda q_* + (1-\lambda)q_{**}) \leq \lambda C(q_*) + (1-\lambda)C(q_{**})$$

The stipulation of a particular hyperplane defined with $a_i > 0$ guarantees that costs can be transray convex regardless of their overall returns to scale and their nonconvexity along rays through the origin. Hence the terminology.

This attribute's economic content is clarified by taking $q_* = (q_1, 0)$ and $q_{**} = (0, q_2)$. While $\lambda q_* + (1-\lambda)q_{**}$ involves less of each good, it involves simultaneous production of both goods. Thus (7) implies that the cost savings from economies of scope outweigh the effects of any product-specific increasing returns to scale. For a twice-differentiable cost function over two outputs, (7) holds if and only if $a_1^2 C_{22} + a_2^2 C_{11} - 2a_1 a_2 C_{12} \geq 0$ on the hyperplane. It suffices for this that the incremental cost of product i be convex in q_i (which implies $IAIC_i$) and that $C_{12} \leq 0$ (which is termed weak cost complementarity and implies weak economies of scope).

This specialization of (4) is transray convex for any $a > 0$, and has economies of scope, $IAIC$ of each product, and U-shaped RAC curves:

$$C(q_1, q_2) = F + (k_1 q_1^{b_1} + k_2 q_2^{b_2})^{\alpha}$$

where $0 \leq \alpha \leq 1$, $\alpha b_i \geq 1$, $k_i > 0$, $q_i > 0$, $F > 0$, $q_1 + q_2 > 0$; and $C(0,0) = 0$.

V. Multiproduct Competitive Industries[6]

All of the preceding cost concepts are needed to understand multiproduct competitive industries. Suppose that goods 1 and 2 are linked by economies of scope. In competitive equilibrium only one of the goods can be manufactured by single product firms for, otherwise, cost-saving mergers would be profitable. Further, zero profit with marginal cost pricing requires that each firm operate where $S = 1$, at the minimum of its RAC curve. Let M denote the set of such minima.

The following analysis of the viability of a competitive multiproduct market structure is the natural generalization of the single product procedure. Find a point, q^*, on M such that if prices were set equal to the marginal costs at q^*, $C_i(q^*)$, then market demand would be a multiple of q^*. If that multiple were large, there may be a competitive market comprised of a large number of price-taking, profit-maximizing firms each earning zero profit at q^*.

In general, the infirmity of this procedure is that q^* may not maximize the profit of a competitive firm. For example, dropping product 1, priced at $C_1(q^*)$, changes the firm's profit by $[C(q_1^*, q_2^*) - C(0, q_2^*)] -$

[5]This concept originated in an early draft of Baumol (1977).

[6]Most of the material presented here is drawn from Panzar and the author (1978).

$q_i^* C_i(q^*)$. This is positive if, at $q^*, S_i > 1$ and $DAIC_i$ holds. Thus a necessary condition for competitive equilibrium is $S_i \leq 1$ for all products i sold by the firm. With economies of scope, (6) shows that there must be $IAIC_i$, for some i.

Thus, information on ray average costs and overall scale economies is insufficient to analyze multiproduct perfect competition. Points on M with increasing returns in any product line (see Figure 1) cannot be produced in competitive equilibrium. If our conceptual procedure only uncovers values of q^* with $DAIC$, then competitive equilibrium either requires firms of several types, or it does not exist at all (both cases can occur).

There is a condition similar to transray convexity that obviates these difficulties. If the cost function is strictly convex on each line segment connecting two points in M, then (given some technicalities) the requisite q^* does exist in M, it does maximize the firm's profit, and the conceptual procedure does reliably test the viability of perfect competition

VI. Multiproduct Natural Monoply[7]

Overall scale economies and ray average costs are also insufficiently powerful for analysis of the monopoly end of the spectrum of multiproduct market forms. In fact, globally increasing returns to scale, even together with economies of scope, do not imply that costs have the natural monopoly property of subadditivity:[8] $C(q_*) + C(q_{**}) > C(q_* + q_{**})$. However, globally increasing returns to scale in a product line do require monopoly of that product for industry cost minimization. Further, $DAIC$ in each product line, together with economies of scope, *does* imply subadditivity.

A natural monopoly firm may be unsustainable; that is, vulnerable to market incursions (that raise total industry cost) by entrants who assess potential profits assuming the incumbent firm will maintain its prices. Consider a natural monopolist charged with producing two goods with $DAIC$ that are demand substitutes. It must forgo some scale economies in each product to protect the scale of manufacture of the substitute. An entrant, however, may profit from undercutting one of the natural monopolist's prices and fully exploiting that product's scale economies, even while, by specializing, losing the economies of scope enjoyed by the monopolist.

Transray convexity of costs and overall scale economies guarantee that economies of scope outweigh product-specific scale effects sufficiently to prevent profitability of such entry. Further, these conditions imply cost subadditivity and (with some ancillary conditions) the sustainability of the natural monopoly. Moreover, under these conditions, the Ramsey optimal set of products and prices is guaranteed to be sustainable.

Thus, for both perfect competition and natural monopoly, new complexities arise when economies of scope lead to multiproduct firms. In each case, some insight is gained from analyzing product-specific scale effects in addition to those overall. And yet, at both extremes of the spectrum of market structures, the concept of transray convexity eliminates the complexities and permits analyses analogous to those of single product industries.

VII. Towards a Unifying Framework

Analyses relating multiproduct technology to other market forms must be based on a vision of industry equilibrium. One approach, fruitfully explored by Baumol and Dietrich Fischer, is to hypothesize directly that market structure will tend, via firm mergers and fissions, towards that which minimizes the total cost of the industry's total output. However, industry cost minimization may

[7]The material in this section is drawn from Baumol (1977), Panzar and the author (1977a), and from Baumol, Elizabeth Bailey, and the author. Critical precursors to this work were Gerry Faulhaber, and Faulhaber and Edward Zajac. The term sustainability was coined in a study sponsored at New York University by the Office of Science Information Services of the National Science Foundation.

[8]Andrew McLennan (private correspondence) was the first to construct an example verifying this.

require that *multiproduct* firms realign their output mixes in a manner unachievable by mergers or fissions. Nonetheless, the following treatment shows that industry cost minimization can be enforced by free entry.

A *feasible industry configuration* over the product set N is m firms producing output vectors $q^1, \ldots, q^m \epsilon R^n_+$, at prices $p \epsilon R^n_{++}$, such that $\Sigma q^j = Q(p)$, where $Q(\cdot)$ is the vector valued market-demand function, and such that $p \cdot q^j \geq C(q^j)$. An industry configuration is *sustainable* if $p^e \cdot q^e - C(q^e) \leq 0$ for p^e, $q^e \epsilon R^n_+$, $p^e \leq p$, and $q^e \leq Q(p^*)$ where $p^*_j = \min(p^e_j, p_j)$.

Sustainability analysis assumes frictionless free entry using the same technology and input markets as the industry's firms. An entry plan consists of prices p^e not exceeding existing prices, and quantities q^e not exceeding market demands at the lowest available prices p^*. An industry configuration is sustainable if no such marketing plan can earn positive profit. This definition generalizes that of sustainable natural monopoly prices. Further, industry competitive equilibria are sustainable feasible industry configurations.[9] Thus, the new definition does provide a unifying framework.

PROPOSITION: *A feasible industry configuration, p, $q^1, \ldots, q^m$, is unsustainable unless it minimizes the total industry cost of producing $\Sigma q^j \equiv q$.*

PROOF:

Suppose there exist $\hat{q}^1, \ldots, \hat{q}^k$ such that $\Sigma \hat{q}^j = q^i$ and $\Sigma C(\hat{q}^j) < \Sigma C(q^i)$. Then $\Sigma(\hat{q}^j \cdot p - C(\hat{q}^j)) = \Sigma(q^i \cdot p) - \Sigma C(\hat{q}^j) > \Sigma(q^i \cdot p) - \Sigma C(q^i) = \Sigma(q^i \cdot p - C(q^i)) \geq 0$. Hence, for some j, $\hat{q}^j \cdot p - C(\hat{q}^j) > 0$. Setting $p^e = p$ and $q^e = \hat{q}^j \leq \Sigma q^i = Q(p)$ shows that unsustainability follows.

Thus, free entry, as described here, assures that industry costs are minimized in equilibrium. Consequently, the idea of sustainable feasible industry configurations provides a general framework for the analysis of the connections between multiproduct technology and market structure.

[9] Further, price-Nash monopolistic competition equilibrium among zero profit single product firms is sustainable. However, sustainability may be too strong a requirement of homogeneous product oligopolistic equilibria since here it implies marginal cost equal to price.

REFERENCES

W. J. Baumol, "Scale Economies, Average Cost and the Profitability of Marginal-Cost Pricing," in Ronald Grieson, ed., *Essays in Urban Economics and Public Finance in Honor of William S. Vickrey*, Lexington 1976.

——, "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry," *Amer. Econ. Rev.*, Dec. 1977, *67*, 809–22.

——, **E. E. Bailey, and R. D. Willig**, "Weak Invisible Hand Theorems on the Sustainability of Prices in a Multiproduct Natural Monopoly," *Amer. Econ. Rev.*, June 1977, *67*, 350–65.

—— **and Y. Braunstein**, "Empirical Study of Scale Economies and Production Complementarity; The Case of Journal Publication," *J. Polit. Econ.*, Oct. 1977, *85*, 1037 49.

—— **and D. Fischer**, "Cost Minimizing Number of Firms and Determination of Industry Structure," *Quart. J. Econ.*, Aug. 1978, *20*, 439–67.

G. R. Faulhaber, "Cross-Subsidization: Pricing in Public Enterprises," *Amer. Econ. Rev.*, Dec. 1975, *65*, 966–77.

—— **and E. E. Zajac**, "Some Thoughts on Cross-Subsidization in Regulated Industries," econ. disc. paper no. 48, Bell Laboratories 1976.

J. Panzar and R. D. Willig, "Economies of Scale and Economies of Scope in Multi-Output Production," econ. disc. paper, Bell Laboratories 1975.

—— **and** ——, (1977a) "Free Entry and the Sustainability of Natural Monopoly," *Bell J. Econ.*, Spring 1977, *8*, 1–22.

—— **and** ——, (1977b) "Economies of Scale in Multi-Output Production," *Quart J. Econ.*, Aug. 1977, *91*, 481–94.

—— **and** ——, Economies of Scope, Product Specific Economies of Scale, and

Multiproduct Competitive Industries," unpublished paper, Bell Laboratories 1978.
F. M. **Scherer et al.**, *The Economics of Multiplant Operation: An International Comparison Study*, Cambridge 1975.
R. Schmalensee, "Entry Deterrence in the Ready-to-Eat Breakfast Cereal Industry," *Bell J. Econ.*, Autumn 1978, 9, 305–27.

The Performance of Government in Energy Regulations

By WALTER J. MEAD*

An evaluation of government performance in energy regulation requires specification of a standard. Here the record of regulation will be evaluated in terms of optimum resource conservation, defined as the process of maximizing the present value of all resources at any point of time.

It is possible to distinguish two general paths by which resource conservation may be attained: free market allocation; or allocation by direct government regulations, the use of taxation, or the use of subsidies. There is wide agreement among economists that the existence of net externalities creates market failures and suboptimal resource allocation. The presence of externalities therefore has been used to rationalize government intervention in order to correct for market failures. The implicit assumption is that such regulation will in turn be economically efficient.

An influential body of opinion argues that the power of government should be used to enforce energy conservation as a means of resolving the "energy crisis." However, most of these arguments are based on naive definitions of conservation, meaning "use less" or "save energy" as if all other resources were of zero value. Concepts of this kind are found repeatedly in the Energy Policy Project report of the Ford Foundation.

The Carter "National Energy Plan" proposes a major expansion of government regulation, taxation, and subsidies. Before Congress legislates new government intervention in the energy sector, would it not be wise to evaluate the extensive record of past government intervention? In this paper five major areas of government intervention over the past half-century will be briefly reviewed.

I. A Review of *U.S.* Energy Policy

More than a half-century ago Congress introduced percentage depletion allowance tax treatment for oil and gas production Subsequently, provision was made for expensing of intangible drilling costs. The initial effect of these two policies was to increase the after-tax rate of return on investments in oil and gas exploration and production. These tax subsidies led to increased capital flows into exploration. Consequently, new reserves were found and production was stimulated. But increased production led to lower oil prices and established the historic *U.S.* low-price policy for energy. This in turn led consumers to treat oil and gas as cheap commodities and to consume these nonrenewable resources excessively.

Legislation in 1975 removed the benefits of percentage depletion allowances for integrated oil companies only. For the nearly 10,000 independent oil and gas producers, the depletion allowance has been retained but the tax benefits have been reduced. The benefits of expensing intangible costs have been retained in total.

These two tax policies have probably been the most important items of government interference in the petroleum industry. In the absence of these artificial stimulants the market would have delayed production. Thus they contributed directly to the energy crisis of the 1970's and in general are counterproductive of a conservation goal.

Prorationing was authorized by the federal government in the 1930's and implemented subsequently by state governments. Stephen

*Professor of economics, University of California-Santa Barbara.

McDonald has shown that *MER*-type prorationing rules, with their depth-acreage allowable schedules and well spacing regulations, are purely arbitrary and are economically inefficient as a solution to the common property resource externality. He has proposed an efficient solution in the form of mandatory unitization. Prorationing also included market-demand restrictions. This form of government intervention has been inoperative since 1972. For nearly four decades it created idle capacity and hence resource waste in the form of excessive investment in oil exploration and production. Thus, both *MER* and market-demand types of prorationing are inefficient and are counterproductive of a conservation goal.

Import quotas, introduced by President Eisenhower in 1959, were partly a consequence of market-demand prorationing. A protected domestic market is inconsistent with free trade. The effects of import quotas were to restrict the supply of imported oil, to increase domestic prices, and to artificially stimulate additional domestic production of this nonrenewable resource. Thus, *U.S.* import quotas distorted the pattern of oil production worldwide and led to excessive production from rapidly declining *U.S.* resources, contributing directly to the energy crisis of the 1970's. An efficient solution to the dependence problem was not introduced until 1975 when the Strategic Petroleum Reserve was authorized by Congress.

Price controls on natural gas were introduced in 1954 and those on crude oil in 1971. In both cases, controls were administered in favor of artifically low prices. In the case of natural gas, low prices led to high demand and low supply thus creating the usual shortage. Consumers who received natural gas allocations consumed it lavishly as a cheap commodity. In this respect the conservation objective has been thwarted. Part of the supply-demand gap has been filled by a close substitute, oil. This has led to increased imports and consequent dependence and balance-of-payments problems.

The effects of oil price controls are not as clear as in natural gas. Charles Phelps and Rodney Smith studied price controls and other regulations in the oil industry and concluded that "the controls have not reduced the prices of refined products" (p. v). Another study recently completed by Robert Deacon and myself tested two hypotheses involving comparisons of domestic and foreign wholesale gasoline prices between 1971 and 1977. Both tests indicated that *U.S.* price controls effectively lowered the price of gasoline through mid-1976. However, from that point through 1977, the evidence failed to support the hypothesis that price controls lowered gasoline prices.

If gasoline prices are not currently below levels which would be attained in the absence of price controls, then demand is unaffected by price controls. However, producers of crude oil receive artificially low prices under the price control system. Phelps and Smith pointed out that, "While the price controls on crude oil did not influence product prices, they did transfer profits within the petroleum industry" (p. vii). Wealth was transferred from crude oil producers to refiners, and from one refiner class to another. If domestic crude oil supply elasticity is greater than zero, then domestic suppliers are artificially restrained by the control system and imports are artificially stimulated, again leading to higher imports, balance of payments, and dependency problems.

Clearly, natural gas price controls have set prices below equilibrium levels. Further, evidence supports the point that oil price controls had similar effects through mid-1976. At lower prices consumers demand more. In the case of petroleum, there is an open-ended supply in the form of imports. If consumption is artificially stimulated, then conservation goals are not attained. In the case of oil and gas price controls there is further resource misallocation in the high cost of price control administration, both within the government and on the part of complying industry.

Arguments against eliminating price controls and allowing the market to allocate scarce oil and gas resources take three forms: 1) World oil prices are alleged to be monopoly prices set by an Organization of Petroleum Exporting Countries (*OPEC*) cartel. Support-

ers of price control argue that *U.S.* oil and gas prices should not be permitted to rise to such monopoly price levels. 2) Market-clearing prices would unfairly impact on the poor. 3) Market-determined prices would confer windfall profits on oil and gas producers.

The cartel rationale in support of price controls was recently articulated by Paul Davidson. But the evidence for the cartel thesis is mixed. On the one hand, the dominant "firm" in the alleged cartel is clearly Saudi Arabia. In the years from the strong crude oil market in 1973, to the relatively weak markets in 1975–77, Saudi Arabia expanded its market share from 24.2 to 30.4 percent of *OPEC* production. This evidence is inconsistent with either a fixed market shares or a dominant-firm price-leadership model of oligopoly behavior. On the other hand, *OPEC* output, in the aggregate, is consistent with cartel behavior. The *OPEC* share of world crude production declined from 55.5 to 52.5 percent during the increasingly weak markets from 1973 to 1977.

Some new research by Ali D. Johany has shown that crude oil price movements from approximately $3 per barrel in the early 1970's to approximately $12.50 per barrel in 1974 are rational in terms of individual oil producing countries maximizing the present value of their resources. By joining property rights theory to capital theory, Johany has carried our understanding of optimal crude oil prices beyond recent work by William Nordhaus and Robert Pindyck.

Johany pointed out that during the 1950's and 1960's there was a progressive awareness on the part of international oil companies holding oil concessions in the Middle East that their property rights were in jeopardy. Nationalization, or its euphemism, "participation," was the wave of the future. Fear of loss of property rights caused international oil companies to accelerate their foreign production. From 1950 through 1970 the compound annual growth rate in oil production from the Middle East was 10.9 percent. From 1970 through 1973, the growth rate was 15.0 percent.

Reflecting these output increases, world crude oil prices were relatively stable during the two decades from 1950 through 1970. Output increases were matched by worldwide growth in demand with only modest increases in nominal prices.

By the end of 1973, however, host countries were in complete control of output within their borders. Given firmly established property rights and lower discount rates, one would expect reduced output growth rates and sharply higher prices. The record shows that from 1973 to 1977 Middle Eastern oil output increased at a compound annual rate of only 0.7 percent. As a consequence, crude oil prices in world markets rose sharply from 1970 to date, a fact which can be explained without the aid of a cartel theory.

Johany examined the opportunity cost of capital for Saudi Arabia and concluded that for large sums of money, the *U.S.* government Treasury Bill market yielding 8 percent appeared to represent the most attractive alternative to leaving oil in the ground. Adjusting for inflation, the real opportunity cost was judged to be 1 percent. If one believes that the real price of oil fifty years hence will be approximately $21 per barrel, as determined by the cost of oil substitutes, then present values are rational. If so, the *U.S* crude oil price controls cannot be justified in terms of a cartel theory.

Oil and natural gas price controls as methods of subsidizing the poor are haphazard tools. If additional public subsidies to the poor are warranted, then a direct approach through the negative income tax device would be a more efficient means.

Finally, decontrol would confer windfall profits on oil and gas resource owners and lease holders. But federal and state governments own most of the known and probable future productive oil and gas resources. That part of the windfall accruing to lease holders would be subject to income taxes. In 1975, the percentage depletion allowance was totally removed for all integrated oil companies. Therefore, combined federal and state income taxes would capture close to half of any windfall gains which accrue to integrated producers. For future leases issued after decontrol, the auction bidding system would totally eliminate windfalls due to decontrol.

A final element of government policy toward energy is the process used in auctioning leases. In 1978, Congress enacted a major overhaul in legislation governing outer continental shelf (*OCS*) oil and gas leasing. Under prior leasing procedures, tracts were leased after oral auction cash bonus bidding. The winning bidder obtained the right to explore for oil and gas resources. If production were undertaken, a one-sixth royalty had to be paid. However, motivated by a belief that competition under the present system is inadequate and government is not receiving "fair market value" for its resources, Congress mandated the use of bidding systems other than cash bonus bidding.

However, recent research has shown that competition for *OCS* oil and gas leases is intense and as a result the government has received more than a fair market value for its leases (R. O. Jones, the author, and P. E. Sorensen). Preliminary findings indicate that the nominal internal rate of return generated by lessees on 839 leases issued between 1954 and 1962 was 9.5 percent *before taxes*. This is a subnormal profit level and indicates that the winning bidders bid too much for their leases. Economic theory, supported by this analysis of the bonus bidding record, indicates that under the proposed bidding systems, the goals of resource conservation will be sacrificed.

II. The Political Economy of Public Policy

This short and necessarily superficial review of major past and emerging energy policies indicates that government intervention has been counterproductive with respect to resource conservation. The reasons are to be found in that ancient discipline known as political economy.

The main concerns of a politician are to get elected and to continue in office. These concerns require that politicians individually and collectively respond to dominant organized pressures brought to bear on them. Pressures from the oil industry obtained and then sustained tax subsidies and market-demand prorationing. The coal industry joined with the domestic oil industry to obtain import quotas. The political power of the oil industry has declined since the early 1970's, to be replaced by environmentalists and consumerists. These groups, together with organized labor, appear to sustain price controls. Concerns for optimum resource allocation are not primary concerns of politicians.

Second, economists advanced the externalities (market failure) concept. Informed laymen and politicians have embraced the concept. But political scientists have been slow to point out its counterpart in the political framework. If congressmen do not bear the full cost of the positions which they take and the consequent legislation which is enacted, then political market failure occurs. Where benefits of legislation are concentrated (the beneficial interests know who they are and are grateful) and costs are dispersed (those who bear the costs are not well informed and the cost per person is low) the net political externality may be significant.

Third, the legislative process is a compromising process. Economists may agree on the character and the extent of a tax subsidy or regulation necessary to correct an externality. But any agreed upon correction must pass through the political process where hearings are held and interest groups have a right to bring pressure to bear on their elected representatives. Political scientist Daniel Ogden pointed out that ". . . national policy is made through a system of power clusters," and further, . . . " administrative agencies jealously guard their subject matter 'turf.' They yield jurisdication only after a major struggle and only in the face of overwhelming political force." Individual congressmen will approve only what is acceptable by the dominant pressure groups to which they must be responsive. What economists believe to be an appropriate correction for an externality is not what is likely to emerge from the political process.

Fourth, whatever emerges in legislation must then be administered. Another political scientists, Marver Bernstein, wrote that "the history of (regulatory) commissions indicates that they may have survived to the extent that they have served the interests of the regulated groups" (p. 73).

Finally, the presence of a net externality is not a sufficient justification for government intervention. The costs of correction, including the costs added in the legislative compromise process and actual administration accommodations referred to above, must be less than the cost of the net externality to be corrected. Failure to meet this test will lead to even greater resource misallocation.

One might assume that with the declining political power of the oil industry in the last decade, future energy policy will be legislated in the national interest. However, the only change is that the power of one interest group has been displaced by others. The structure of public policy formation as outlined above is unchanged.

President Carter has called for a "comprehensive national energy policy" and his "first principle" asserts that "we can have an effective and comprehensive energy policy only if the Federal government takes responsibility for it. . . ." (Office of the President, p. 1). The record of past energy policy does not lead one to be confident that more intervention will improve resource allocation. An alternative national energy policy would be to let the market allocate scarce resources.

REFERENCES

Marver Bernstein, *Regulating Business by Independent Commissions*, Princeton 1975.

P. Davidson, "Beware the Modern Tripoli Pirates of Natural Gas," *Los Angeles Times*, Aug. 6, 1978.

R. T. Deacon and W. J. Mead, "Price Controls and International Petroleum Product Prices," final report to the Federal Energy Administration, June 16, 1978.

A. D. Johany, "OPEC is Not a Cartel: A Property Rights Explanation of the Rise in Crude Oil Prices," unpublished doctoral dissertation, Univ. California-Santa Barbara, June 1978.

R. O. Jones, W. J. Mead, and P. E. Sorensen, "Economic Issues in Oil Shale Leasing Policy," *Oil Shale Symposium*, forthcoming.

Stephen L. McDonald, *Petroleum Conservation in the United States*, Baltimore 1971.

W. D. Nordhaus, "The Allocation of Energy Resources," *Brookings Papers*, Washington 1973, *3*, 529–70.

D. M. Ogden, Jr., "Protecting the Energy Turf The Department of Energy Organization Act," *Natural Resources J.*, Oct. 1978, *18*.

Charles E. Phelps and Rodney T. Smith, "Petroleum Regulation: The False Dilemma of Decontrol," Rand Corp. R-1951-RC, Jan 1977.

R. S. Pindyck, "Gains to Producers from the Cartelization of Exhaustible Resources," *Rev. Econ. Statist.*, May 1978, *60*, 238–51.

Energy Policy Project of the Ford Foundation, *A Time to Choose*, Cambridge, Mass. 1974.

Office of the President, "Detailed Fact Sheet, The President's Energy Program," Washington, Apr. 20, 1977.

The Role of the Government in Subsidizing Solar Energy

By MICHAEL D. YOKELL*

This paper discusses the economic rationale for a federal solar energy subsidy program, the general type of program which is required, and methods for determining the proper funding level for each program. Nonsubsidy programs, such as the provision of federal standards for solar manufacturers, are not discussed.

I. Should the Federal Government Subsidize Solar Energy?

The federal government should subsidize solar energy for both "first best" and "second best" reasons. First best policies are those which could increase social welfare even if the economy was functioning according to the normative standard of perfect competition. Second best policies are those which are designed to increase social welfare in an imperfectly competitive economy (see E. J. Mishan, p. 202). First best reasons include: 1) innovation often involves a public good externality because the social returns from innovation cannot be entirely captured by an innovator; 2) investors in the innovation process are usually risk averse when investing in new technologies. By pooling the risk over a far larger number of individuals, society should willingly support a higher level of investment in risky innovations than private firms or individuals would; 3) private purchasers of new technology are averse to risk of product failure, improper installation, etc. The private willingness to purchase an innovative product would be raised to optimal levels if risk could be pooled; 4) capital market imperfections can prevent a purchaser from taking advantage of potential life cycle savings offered by solar technologies with high initial costs.

Second best reasons for subsidizing solar energy include: 1) large, partially unjustified subsidies to conventional energy technologies currently cause them to be overused relative to solar energy technologies; 2) average pricing arrangements for petroleum, natural gas, and electricity cause them to be overused relative to marginally priced solar energy technologies; 3) large, partially unpriced environmental external diseconomies result from the production and consumption of energy from conventional sources. Conventional energy sources are overused relative to energy from solar technologies, which generally involve significantly less environmental impact.

For each second best problem, there is a first best remedy applicable, as noted below. Unfortunately, these remedies do not appear to be politically viable at present. Generally speaking, this is because first best solutions involve removing existing subsidies which have developed politically powerful constituencies. Offsetting subsidies may be politically viable because they too have or are developing constituencies. Each of these reasons for subsidizing solar energy will now be discussed briefly.

An innovation in production is not merely the invention of a new product or process, but requires that the product or process actively penetrate the appropriate market. Innovations per se are thus not patentable. Patent laws do not generally afford great protection to the inventor in a field of great technical ferment, both because generic ideas are not patentable and because relatively minor technical changes on a basic technological theme often result in the award of new patents. Thus

*Senior economist, Solar Energy Research Institute. The views expressed here are my own and do not necessarily represent those of the Solar Energy Research Institute, its operator, the Midwest Research Institute, Inc., or the U.S. Department of Energy. Helpful comments have been received from Dennis Costello, Sam Flaim, Bert Mason, Marty Murphy, and Lewis Perelman. A longer version of this paper is available upon request.

a firm which pioneers an innovation may expect to share the profits from the innovation with other firms in the same field. Moreover, an innovation can result in significant benefits to consumers, which they, and not the innovator, capture as consumers' surplus. Thus on both the production and consumption side, an innovator cannot necessarily capture the full social benefits of his innovation and is, therefore, quite unlikely to invest a socially optimal amount in its development.

Even if an innovator were assured of capturing the full social benefits available from a particular innovation, there is no advance assurance that these will be positive. The private investor, even a large one, is often risk averse, and thus, innovation does not take place unless high returns are foreseen by the investor. Because the loss from any unsuccessful innovation is a much smaller fraction of society's assets than it is of even the largest firms, society may well be significantly less risk averse to this loss than any private firm. In this case, again, less than socially optimal amounts are invested in innovation. By subsidizing the process, the federal government can raise the level of innovation to socially appropriate levels.

A very similar phenomenon can be seen on the consumer's side of the market. Any individual innovative product is likely to provide only a relatively small increase in consumer's surplus for the typical consumer. Yet if the product fails, the attendant loss of consumer's surplus may be large, relative to the consumer's total resources. Even if the expected returns are positive, the typical consumer will be risk averse, and therefore consume fewer innovative products than he would if the loss from product failure were small enough to significantly reduce his risk aversion. (This, of course, assumes a particular type of utility function for the consumer, but one which is commonly used.) This reduction in potential loss can be achieved by spreading such losses over many consumers through a risk-pooling mechanism such as insurance. The increase in social welfare attendant upon product failure risk pooling is sufficient justification for this type of arrangement, but it does not necessarily require federal subsidies. Private insurance markets, usually in the form of product service contracts, often develop to provide the essential risk pooling. However, before an actuarial record has been developed (and thus uncertainty is great) private insurers may be unwilling to enter the insurance market. Thus, if solar technologies provide positive net social benefits, temporary federal support for solar equipment warranties is desirable.

Solar technologies (with the exception of biomass) do not require much fuel. As a fraction of life cycle costs, "up front" costs are higher for solar than for most other energy technologies. For the end user, financing arrangements are therefore often more crucial for solar than for other energy technologies. Suppose we are considering a solar technology whose present discounted costs are lower than those of a competing conventional system (i.e., the solar system is "economic" according to the conventional definition). In today's residential loan market, the solar user may have great difficulty financing the incremental costs of a solar system relative to a conventional system. In retrofit applications, loans are often unavailable, or available at home improvement rather than lower mortgage interest rates. For both the business and residential user, the problem is the lender's use of cash flow rather than economic criteria for evaluating loan proposals. Economically feasible solar systems may have long discounted payback times, exceeding the length of typical business loans. If the user has a cash flow problem, which he often does, a solar system becomes an unattractive option. Thus in both the residential and business cases, federal intervention is called for to improve the functioning of the capital market and insure that lending institutions encourage rather than discourage economically feasible solar energy systems.

Before discussing second best reasons for subsidizing solar energy, it might be useful to ask whether innovations in solar energy are different from other technologies. The answer is no. Thus, for the first subsidy argument to be valid, the "traditional" view that private firms invest a less than socially optimal amount in innovative activity must hold (but see Jack Hirshleifer, and Morton Kamien

and Nancy Schwartz). Second best arguments are more directly concerned with solar energy technologies.

A recent report by Bruce Cone et al. catalogues federal subsidies to energy production and estimates their magnitude. The total subsidy was estimated to range between 123.6 and 133.7 billion undiscounted dollars since 1918. Some of these subsidies can be regarded as first best attempts to inprove the functioning of the private energy market, but many of them have little justification other than political expediency. Many examples come to mind, particularly from the petroleum industry, which received 60 percent of the subsidies (not including those to natural gas). Most of these subsidies result in a market price below that which would exist under perfect competition. In this situation, solar energy systems are competing against artifically low-priced conventional energy systems. A first best solution is to remove the unjustified portion of subsidies from conventional energy sources. This currently appears politically impossible, although progress has been made by eliminating the oil depletion allowance. A second best alternative may be to provide federal subsidies for solar systems to compensate for subsidies to conventional fuels, and thus prevent distortion in interfuel competition. While there is no general theoretical proof that an intervention of this type results in a net increase in social welfare (see Richard Lipsey and Kelvin Lancaster), it is likely to do so in this particular case.

In cases where direct solar end use competes with electricity or natural gas distributed by utilities, a marginally priced good is competing with an average-priced good. A similar situation occurs in the market for refined petroleum products under the "entitlements" program. Residential solar heating and hot water systems competing with electric resistance heating systems (now accounting for about 50 percent of the new market) are a case in point. The end user sees the true marginal cost of the solar system, but sees only an average price for the fuel component of the electric resistance heating system. A first best solution to this price distortion problem would be to require marginal cost pricing by utilities and simultaneously tax away any windfall profits which the utility would capture from this pricing mechanism. Since marginal cost pricing has been implemented in only a handful of the nation's utilities, a second best solution seems to be called for, at least in the near term.

Conventional energy sources are responsible for significant external diseconomies. While there has been no definitive work quantifying in dollars the health, property, crop, etc. damages due directly and indirectly to conventional energy related pollution, very rough calculations indicate they are large. On the other hand, solar energy systems generally have little if any direct pollution associated with them. Indirect pollution created by solar energy systems is currently under study by the author. Because pollution is generally unpriced, pollution-intensive products are underpriced and overused. According to the well-known argument (see William Baumol and Wallace Oates, ch. 3), a first best solution to this problem would be the provision of optimal taxes per unit of pollution, rebated to the public on a per capita basis. Under this solution, pollution-intensive energy systems would be automatically penalized relative to less polluting solar systems.

II. Federal Subsidies For Solar Energy

This section summarizes the policy instruments for subsidizing solar energy. A "policy instrument" can be defined as a generic program. A "program" is a specific instrument whose funding level may be varied independently. This section does not present or discuss specific solar programs. In the analysis of any specific program the administrative costs of the program, as well as the costs of any unintended effects, must be weighed carefully against the main benefits of the program.

A number of federal policy instruments and programs are appropriate for subsidizing the development and application of solar energy technologies. How many should be simultaneously undertaken? Each of the four first best problems outlined requires a separate policy instrument if the effect of each

solution is to be independently targetable. The second best problems, however, all join to create one fundamental problem: conventional sources of energy are underpriced relative to solar sources. One policy instrument (which could be applied independently in several programs for different solar sources to reflect the extent to which their conventional competitors are underpriced) is sufficient to correct all three second best problems. Thus, five policy instruments are required to provide generically optimal subsidies for solar energy. If additional policy objectives are added to a solar program, then additional policy instruments are required to independently target the level at which these objectives are met.

The broad outlines of an economically optimal and socially acceptable solar subsidy program may now be presented. First, a major program to compensate for the underpricing of conventional energy sources is required. Subsidies which operate on the demand side rather the supply side of the market are preferable because their benefits are spread more broadly. The major options for providing such subsidies may be ranked in order of increasing progressivity as follows: tax deductions, tax credits, tax rebates, and below-market-rate loans for solar end users. If the purported federal objective of progressive income transfers is adhered to, then this should be the order of increasing preference. Second, a major program to reduce the end user's perception of risk is warranted. Here it may be wise to provide different programs for different types of end users using different technologies. For the homeowner, federal cost sharing of warranties on solar systems is likely to be the best program. Industrial end users of large amounts of energy would probably be more influenced by major federal demonstration programs. Commercial users probably stand somewhere in between. Third, a major program is required to reduce the risk of innovation and compensate for the difficulty firms have capturing the full benefits of innovation. Major federal research and development programs are clearly warranted. In special cases, federal equity participation may also be warranted.

III. The Proper Balance Among Federal Programs To Encourage Solar Energy

Now that a broad federal solar program has been outlined, the level at which each is conducted must be addressed. In theory, each program should be increased in size until marginal social benefits from the program equal marginal social costs. This condition should be obtained by running federal programs at a level just sufficient to overcome the problems outlined in the previous section. For example, subsidies which are given to solar technologies to lower their cost relative to subsidized conventional alternatives should be just sufficient to compensate for the subsidies which now act to reduce the price of conventional energy. In practice the problem is considerably more complex. Each solar technology competes in only a segment of the energy market; thus, each solar technology should be subsidized at a different level depending on the average subsidy which affects the price of its competitors. Since some technologies compete in more than one market a perfect solution is administratively impossible. Determining the required correction for first best problems is also difficult. For example, suppose we want to compensate for private innovators' inability to capture the full benefits of innovation by establishing a federal research and development program. To do this at the "optimal" level would require us first to determine what the investment behavior of innovators would be if they *could* capture the full benefits of innovation. A research and development subsidy would then be provided which, when added to innovators' private investment in research and developments, would equal that level of investment which would occur under first best conditions. A further complication is the possibility that public investment would affect the level of private investment and the two would not be additive.

In the actual policymaking process, the politically desired level of solar market penetration in specific markets is determined, and then programs are funded to levels which are judged to stimulate that level of market

penetration. Even using this "reverse" methodology, the problem is complex. Market penetration must be established under varying subsidy types and levels.[1] In cases where the proposed subsidy affects costs directly this is difficult enough; in a case like research and development subsidies, where their prior impact on costs is unknown, it is an extraordinary problem.

If market penetration is not the only policy objective, a similar but more complex decision scheme must be employed. Such a scheme was used recently by the Department of Energy's "Solar Working Group" in preparing its review *Solar Energy Research and Development. Program Balance* (see Charles Hitch et al.). In this study, seven choice criteria were used, only one of which was market penetration. Each choice criterion or "value" was assigned a weight for each of three time periods based on the working group's judgement. Different weights were used for each of seven solar technologies, and total "benefit points" assigned to each technology. Finally, marginal benefits resulting from increases in research and development funding levels were calculated for each solar technology. This phase of work relied on judgement as to the likely effect of additional research and development funding on the costs of the various solar technologies.

The Solar Working Group/SRI study considered only program balance within the current research and development effort. A more comprehensive study would use a similar methodology to analyze the program balance within other federal solar incentive programs. Having optimized the balance within programs, the relative merits of each subsidy program need to be analyzed.[2] No comprehensive study has yet done this, but one is surely needed.

[1]The MITRE Corporation and SRI have both developed market penetration models now being used for planning purposes by the Department of Energy.

[2]This two-step procedure assumes that the distribution of benefits among technologies within a subsidy program does not change substantially as funding levels are changed. If this is incorrect, program balance within each program and relative funding levels of various programs must be optimized simultaneously

IV. Conclusion

Designing optimal federal solar subsidy policies and programs requires several phases. First, the nature of the economic problems which might call for subsidies need to be analyzed. Second, the objectives of any general policy must be chosen. Third, policy instruments capable of reaching the proposed objectives must be chosen. Fourth, specific programs for each solar technology must be designed within the general framework of the chosen policy instruments. Finally, the relative funding levels within and among programs must be optimized.

REFERENCES

William Baumol and Wallace Oates, *The Theory of Environmental Policy*, Englewood Cliffs 1975.

Bruce Cone et al., "An Analysis of Federal Incentives Used to Stimulate Energy Production," PNL-2410/UC-50, Battelle Pacific Northwest Laboratories, Mar. 1978.

J. Hirshleifer, "Where Are We in the Theory of Information?," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 31–39.

Charles Hitch et al., *Solar Energy Research and Development: Program Balance*, advance copy, Feb. 1978.

M. Kamien and N. Schwartz, "Market Structure and Innovation: A Survey," *J. Econ. Lit.*, Mar. 1975, *13*, 1–37.

R. G. Lipsey and K. Lancaster, "The General Theory of the Second Best," *Rev. Econ. Stud.*, No. 1, 1956, *24*, 11–32.

M. Yokell et al., "The Environmental Benefits and Costs of Solar Energy: An Economic Approach," SERI/TP-52-074, Solar Energy Res. Instit., Spring 1979.

E. J. Mishan, "A Survey of Welfare Economics 1939–59," in *Surveys of Economic Theory*, Vol. 1, New York 1967.

Another Look at Energy Conservation

By LEE SCHIPPER*

While "moral war" and "national will" have become associated with energy conservation during the past year, there remains only one really important reason to think about how we use energy: it costs less to conserve energy than to produce energy from any new sources. If social and environmental cost and benefits are counted, then the impetus to conserve would be even greater. But what do we really mean by conservation of energy? And why all the controversy?

I. What is Conservation?

Conservation is a response to exogenous changes in relative costs, including possibly external costs. While conservation has many political, social, or environmental connotations, I identify conservation with economic efficiency:

a) Conservation means substituting less costly resources or production factors for energy—mainly capital, but also information, materials, and labor. Capital equipment and processes are thereby changed in the medium and long term. Conservation means minimization of the present value of capital and operating costs (see Figure 1).

b) Conservation means short-term changes in consumer behavior towards a few key energy intensive activities—driving, heating, cooling, hot water use. Existing capital is used. To the consumer the value of the saved energy exceeds the perceived cost of making the change (see Figure 2).

c) Conservation can appear through structural change, either as the cause, or, more likely as the effect. Changes in the market basket of non-energy-intensive goods, changes in the use of land, or long-term changes in behavior and preferences can effect energy use greatly. Other things being equal, the ratio of energy consumed to *GNP* would change.

Of these changes, the first has its greatest effect in the medium term, as existing capital is replaced. More energy can be saved per dollar invested than in retrofit. The second reaction can have a marked effect on existing consumption patterns. The third can lead to enormous drops in the energy requirements of the economy through structural shifts.

In all cases it is resource use and consumer amenity satisfaction that is being optimized, not simply energy use per unit of output Owing to the relative rise in most energy prices, however, economic efficiency will reduce energy intensities in the long run compared to what would have occurred had energy prices continued their historic fall.

Energy elasticities, energy use, and energy conservation must be carefully measured Aggregates like energy/*GNP* are almost worthless for analysis since structure, intensity, and behavior are mixed together. Energy use/output in specific, well-defined processes, factor analysis for well-defined processes, or energy costs/total costs for specific processes are far more suitable as measures of performance. While energy/output for the entire paper industry mixes many processes and outputs, energy use per ton of pulp (which can be further subdivided by pulping process) more accurately measures performance. Similarly buses and autos are not equivalent "processes" except in special cases and should be disaggregated separately from "passenger transport." Ultimately conservation can be measured as reductions in energy/output, energy's factor share, or energy costs/total costs.

II. How Much to Conserve

Contrary to some views that conservation is an all-or-nothing proposition, conservation is a continuous process. In fact, energy/output

*Lawrence Berkeley Laboratory and Royal Swedish Academy of Sciences.

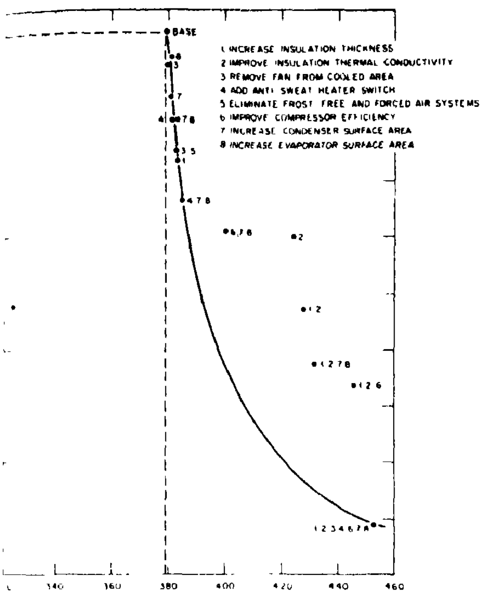


FIGURE 1 ENERGY USE VS. RETAIL PRICE FOR VARIOUS DESIGN CHANGES FOR A 0.45 m^3 (16 ft^3) TOP-FREEZER REFRIGERATOR

gradually has fallen in most manufacturing industries because of technological progress.

It is sometimes said that such conservation implies labor intensive practices or lower productivity. Ernst Berndt and David Wood suggested that capital and energy were complements while the capital-energy bundle and labor were substitutes. Other evidence (see Thomas Long and the author; James Griffin and Paul Gregory) suggested that there was some substitutability between energy and capital; recent investigations (see Savas Özatalay et al.) confirm this directly. Newer, more labor-productive heavy industry requires less energy/product as well as less labor/product than older (see Bo Carlsson). This substitution of capital for labor increased the ratio of energy/labor if labor costs increased while energy prices declined.

A. *Manufacturing*

Industries planning new equipment so as to minimize costs can reduce energy intensities considerably compared to today's uses, at a

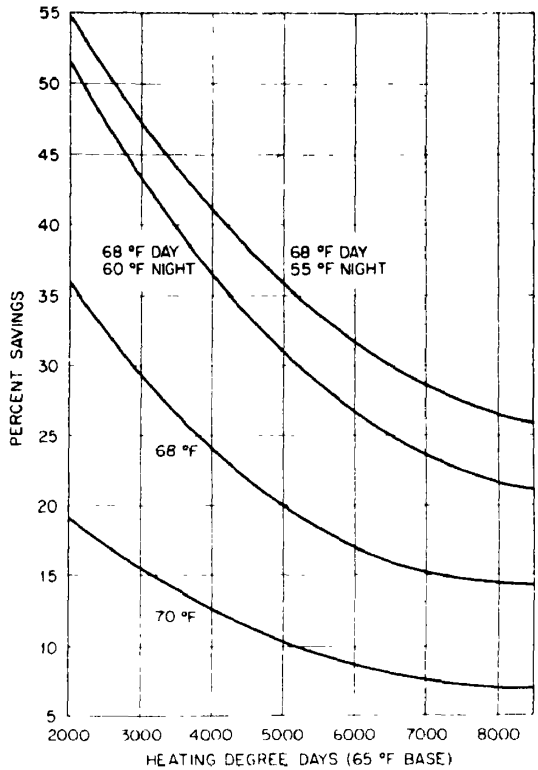


FIGURE 2. PREDICTED ENERGY SAVINGS FOR SEVERAL THERMOSTAT SETTINGS (72° F is the Reference Setting and Night Setback is from 10 P.M. to 6 A.M.)

very small increase in capital. The full cost of capital services/unit of output, while possibly higher now than would have been the case had energy costs continued to fall, will be *lower* than if no measures are taken to reduce energy intensities. Berndt and Wood suggest that this might stimulate the substitution of capital-energy for labor even further, but it is hard to see how this increased capital intensity could remove more than a small part of the energy savings per unit of output "won" by conservation.

Most of the energy savings will be in process heat, not in labor saving motive power; "labor" intensive practices will not return. Heat recovery, more efficient combustion, and process controls will simply replace energy at the margin. It is improper to label these substitutions as deleterious to productivity, since industries that conserve will cut costs. Any loss in productivity growth comes

about through the exogenous rise in the cost of one factor of production, energy.

How far will intensities fall in the future? That depends on the changes in energy prices. Given recent price rises, increases due to marginal cost pricing (= decontrol), and expected inflation in all energy costs, new facilities can be designed to produce raw materials with 20 to 60 percent less energy than existing plant averages. The expected incremental capital cost of these savings will be small, both compared to the value of new industrial equipment and compared to the cost (at the margin) of producing the equivalent amount of new energy supplies. Whether these savings are obtained depends critically on how aware industrial decision makers are of present and future marginal energy costs. It will probably be many years beyond the arrival of marginal cost pricing before the full value of conservation is routinely captured in industrial plant.

International comparisons of energy use, suitably disaggregated and adjusted, hint at how much energy can be saved at various energy prices (see the author and A. Lichtenberg). If production function analyses are suitably disaggregated, these technical hints can be understood in economic terms, as Carlsson or Özatalay et al. show. Thus capital substitutes for energy in steel production in Japan (vs. the United States), while paper, cement, and steel are produced for less energy/ton in higher price energy lands like Japan, Sweden, or Germany, compared with the United States. Table 1 summarizes the *U.S.*-Swedish comparison.

The *U.S.* cement industry is actively modernizing its facilities (see the *New York Times*, Dec. 25, 1977) by substituting larger European dry kilns for smaller, ancient wet kilns. These modernizations improve productivity of labor while cutting energy use by nearly 50 percent per ton of output, in part due to the larger size of new dry kilns. Moreover the new kilns use coal and "eat" the sulfur produced. How fast new kilns replace old depends on overall demand for cement and the point at which the marginal cost of a ton of clinker from an old factory—due mainly to operating costs—exceeds the capital and operating costs of new equipment.

TABLE 1—CONTRASTS IN ENERGY USE: SWEDEN/UNITED STATES RATIOS

	Per Capita Demand	Intensity	Total Energy Use
Autos[a]	0.6	0.6	0.36
Mass transit[b]			
trains, bus	2.9	0.80	2.35
Urban truck[c]	0.95	0.3	0.28
Residential space			
heat[d]	(1.7 x 0.95)	0.5	0.81
Appliances[e]	?	?	0.55
Commercial[f]			
total/sq. ft.	1.3	0.6	0.78
Heavy industry[g]		0.6–0.9	0.92
(physical basis)			
Paper	4.2		
Steel	1.1		
Oil	0.5		
Cement	1.35		
Aluminum	0.5		
Chemicals	0.6		
Light industry[h]	0.67	0.6	0.4
($ value-added)			
Thermal generation			
of electricity[i]	0.3	0.75	0.23

Notes: The Demand column gives the ratio of final demand in Sweden to that in the United States for important energy uses, the Intensity column the relative energy intensities. It can be seen that both factors influence total energy use. In industry, structure in Sweden is more energy demanding than in the United States, but individual energy intensities are lower. Ultimately lower energy intensities in Sweden account for about two-thirds of the difference in per capita energy use.

[a]Swedish 24 MPG-driving cycle uses less energy

[b]Mass transit takes 40% of passenger miles in trips under 20 km in Sweden.

[c]Swedish trucks smaller, more diesels.

[d](Energy/degree day x area) Sweden 9200 degree days vs. 5500 United States degree days.

[e]More, larger United States appliances.

[f]Air conditioning important in United States only.

[g]Sweden more electric intensive due to cheap hydroelectric power. Also Swedish cogeneration.

[h]Space heating significant in Sweden.

[i]Swedish large cogeneration share.

Here the price of energy plays a key role, as Carlsson showed. This time factor is important—energy savings to date in all *U.S.* industry are based on retrofit of older equipment. Future gains will be greater, unless the price of energy suddenly falls, an unlikely event in my own judgment.

Finally, thermodynamic limits to energy

use in production are still a long way from today's intensities. Paper mills that produce all necessary heat and power from wood wastes have been projected; many processes can now valuably sell waste heat rather than discharging it to the environment; electric power and heat can be cogenerated. Technology and relative prices will be the deciding factors, not nature's laws. In all, the link between energy and production, like the link between other factors and production, is extremely flexible in the long run. This is the meaning of energy conservation in the production sector.

B. *Buildings*

In 1970 when real energy prices reached their historical minimum the present value of energy "conservable" in new structures—up to 40 percent of existing heat and cooling—exceeded the capital costs of saving this energy with rates of return of around 10 percent or more. Many institutional barriers hindered the efficient use of resources (see the author et al.). The future will be different. Even before new building codes and techniques appeared consumers began buying insulation in record amounts, spurred by cold winters and higher energy costs. At Lawrence Berkeley Laboratory we estimate that retrofits allow reduction of 20–80 percent of heating loads and 20 percent of cooling loads in homes with rates of return of better than 10 percent. For commercial buildings existing plants can be modified for about a 25 percent saving, again with an attractive rate of return (see the *CONAES* report).

In new structures and equipment the savings are even more dramatic. Compared to today's energy intensities, new refrigerators, new water heaters, and new building shells require fractions (60, 80, 20–70 percent) of today's energy use with incremental investments of the order (10–20, 10, 1–5 percent) of total system costs, giving rates of return exceeding 8 percent. Here as in the industrial sector the effects of price controls, average rather than marginal costing, or subsidies to energy producers (such are the investment tax credit for utilities), are important. In California, for example, present residential natural gas prices (less than \$2/GJ) justify attic insulation and some retrofit wall insulation, as well as clock thermostats, saving 20–40 percent of existing energy use with rates of return greater than 10 percent (see the author and Joel Darmstadter). At parity prices (about \$3/GJ) wall insulation and double windows are profitable in many homes, while at marginal prices (electric heat or synthetic fuels at > \$6/GJ delivered) homes would require very little energy for heating at all. Indeed it is less expensive to eliminate nearly all of the heating load in the "sunny" part of the country than to capture most of the load with solar heat. If electric and fuel prices rise to replacement costs, however, solar water heating will become the least expensive source of this important amenity, and solar space heat should penetrate the heat market somewhat.

C. *Transportation*

In the transportation sector, important energy saving technical changes have been occurring in autos, trucks, and airplanes. On the other hand, modal changes (cars to buses or rail, airplanes to rail and buses, truck freight to rail) seem very unlikely simply as energy conservation measures. This is evidenced by experience in Europe and a multitude of economic and attitude surveys regarding the rise—and fall—of transit. However mass transit, railroads, and buses have other important benefits that far outweigh the energy savings of these modes.

While autos have shrunk, they have also become more efficient technically (i.e., energy consumption per unit of passenger space has fallen). Since 1973 a combination of changes in auto buying habits, shrinking of individual models of cars, and improvements in the efficiency of each model have caused energy intensity to fall by more than 25 percent. Given the drop in real gasoline prices since the initial rise in 1974, consumers might balk at buying smaller cars, and some retrenchment has occurred. Low gasoline prices combined with efficient cars reduce the marginal cost of auto travel considerably,

stimulating the increased use of the car. Clearly the continued "enforcement" of MPG standards depends on society's attitude towards the value of reducing energy use per mile. Higher gasoline prices, or taxes on the MPG or weight of cars, more common in Europe, should be considered to support these goals, especially if short-term conditions force gasoline prices downward.

While the ultimate results of a concerted effort to reduce the energy intensity of auto transport may be dramatic—30+ MPG fleet averages—the changes expected in trucks and airplanes are also important. New powerplants and wind-designs should increase truck effectiveness, as will changes in rules for hauling practices. The European Airbus already shaves total costs and fuel costs in the medium-haul air market, and American manufacturers are reportedly close behind. As the energy intensity of all modes are lowered, the energy-related shifts in modes becomes even less important, an effect worth remembering when the auto and bus are compared.

To summarize the prospects for energy conservation, I have gathered in Table 2 the findings of the Demand and Conservation Panel of the Committee on Nuclear and Alternative Energy Systems (*CONAES*) of the National Academy of Sciences. Shown are the energy intensities of the most important uses of energy, relative to present (1975) practices, for a variety of price and policy futures. The intensities shown represent averages for all systems in place in 2010, including the effect of retrofit on existing plant and structures. In the Panel's judgment these intensities lie near the economic optima for the price futures considered.

TABLE 2—ENERGY INTENSITIES IN 2010

Use:	II	III	IV
Thermal Integrity			
Residential	0.63	0.63	0.76
Commercial	0.42	0.60	0.70
Government, Education	0.35	0.45	0.50
Space Conditioning			
Air	0.66	0.75	0.94
Electric Heat	0.52	0.63	0.90
Gas/Oil Heat	0.72	0.75	0.80
Refrigeration, Freezing	0.58	0.68	0.92
Lighting	0.60	0.70	0.70
Agriculture	0.85	0.85	0.95
Aluminum	0.55	0.63	0.80
Cement	0.60	0.63	0.75
Chemicals (excluding feedstocks)	0.74	0.78	0.84
Construction	0.58	0.65	0.73
Food	0.66	0.76	0.86
Glass	0.69	0.76	0.82
Iron/Steel	0.72	0.76	0.83
Paper	0.64	0.71	0.76
Other	0.57	0.75	0.85
Auto[a]	(37)	(27)	(20)
Lite Truck[a]	(30)	(21)	(16)
Freight Truck	0.6	0.8	0.9
Air Passenger	0.42	0.45	0.5

Notes. 1975 = 1.00. Energy intensities in three *CONAES* scenarios. Average energy prices (use weighted) were 4x, 2x, 1x 1975 levels in Scenarios II, III, and IV, respectively. For details see the *CONAES* report.
[a]Shown in MPG

III. Lessons and Myths

The *CONAES* results and investigations in the United States and in other countries reveal a great degree of technical and economic flexibility in the use of energy. Given changes in relative prices for energy, modest developments in technology and a few key policies regarding standards, other resources will be profitably substituted for energy. The demand for energy depends critically on this elasticity of substitution, as well as upon the income and price elasticities of various energy-intensive amenities, such as driving or space comfort. Many investigations are underway determining both substitution elasticities and the behavioral oriented elasticities.

It is widely alleged that this flexibility is illusory, the counter evidence being regressions of energy use and *GNP* over time or across countries or states. Such work is of little value since structure, price, geography, climate, policy, and the state of the art of energy conservation is omitted. While it is possible that new energy intensive technologies or habits will appear, the rising price of energy makes this unlikely. Indeed new technologies or lifestyles not contemplated by

CONAES might reduce energy intensities. Moreover the most important uses of energy—space conditioning and automobiles—are near saturation, while "new uses," such as calculators, hi-fis, medical care, or hamburger cookers, use insignificant amounts of energy/output and reduce the energy/*GNP* ratio compared to the mix of goods and services prominent in the 1950's.

In my view, then, the link between energy and *GNP* is a flexible one, and that flexibility is being tested now. International comparisons bear these conclusions out (see Table 1). Gains in energy efficiency have been seen in most industrialized countries since 1973, and indeed in the decades previous. Since energy conservation does not threaten labor productivity or lifestyle—we would not conserve energy where that was the case—there seems little need to worry about the "sacrifices" called forth in President Carter's speeches.

IV. The Issue of Price

The key link between energy and the economy appears to be the price of energy. Energy costs play an important role in determination of the optimum balance between energy and other resources. Unfortunately our government and many groups have insisted on a variety of measures that lower the cost of energy to below replacement levels: price controls, tax subsidies, subsidies for new supply systems, and in some cases offsetting subsidies for certain conservation measures.

We must recognize that we are in an era of rising long-run costs of energy. All substitutes for domestic or imported oil and gas will ultimately cost more than these conventional fuels, and the economy must begin adjusting to that situation. Legitimate distributional questions, especially the impact of higher energy costs on the poor, ought to be handled as such, rather than by keeping the price of energy low.

Of course it is often argued that the world price for oil is controlled upward by the *OPEC* cartel. This may be true in the short run, but examination of all alternatives, which are more expensive, suggests that at some time in the near future the market price for world energy supplies, pushed up by growing demand and the high marginal cost of new supplies, will rise above the *OPEC* price, which has stayed nearly constant in real terms for several years. Including environmental costs in the price of energy, not always an easy task analytically or politically, would raise the price of energy even more. Ignoring environmental costs, or subsidizing the new energy sources beyond their normal development stages, would only lead society to overconsume energy (and the environment) relative to other resources.

V. Intervention?

In my view energy prices should represent full social costs of producing and using energy, but this change in pricing policy may be insufficient to bring about changes in the energy system. Supply experts have made it clear (see the *CONAES* study) that massive government intervention in all areas of energy supply will be necessary if energy supplies are to *double* by the year 2010. This intervention will doubtlessly include suppression of environmental standards.

But the government could pay attention to the demand for energy. Many kinds of market failures, related to lack of information, lack of access to capital or lack of influence over the design and operation (or ownership) of energy using facilities have created true economic waste in the buildings sector. Auto MPG standards already on the books have influenced greatly the choice of technologies now employed in automobiles. Industry, on the other hand, is not targeted for end-use regulation, at least as far as energy intensity is concerned.

Is acceleration of the progress of energy conservation politically or socially acceptable? Can we change the maze of building codes or appliance buying habits of consumers and home builders? It seems to me that these difficulties, hard as they are to quantify, must be compared with the enormous difficulties inherent in bringing any new major energy supply options to the market place. Given the environmental uncertainties of *all*

supply options, I would first opt for a minimum of firm, carefully optimized regulations to insure that new buildings, appliances, and homes are built more carefully than in the past. In California, for example, insulation requirements are carefully attuned to the price of energy: no one "forced" to buy insulation is losing money when reasonable interest rates are considered.

Perhaps the most important reason for including key regulations in any policy is a fundamental lack of two other resources: time and certainty. We have made a political judgment that we must hurry to reduce our dependence on imports—dollar for dollar, barrel for barrel, conservation does this faster than new supplies. But both conservation and new supplies have uncertainties in practice. Building codes for example would act to minimize uncertainty over the pace and success of conservation in buildings and, as I have observed in Sweden and California, generally speed up the pace of technological change. Regulations on behavior, on the other hand, whether in the form of maximum temperatures, gasless Sundays, bans on production of certain goods and services, or forms of energy rationing, have no place in any economic system accustomed to at least some degree of freedom of choice, and would hardly contribute to significant long-run energy savings.

Unfortunately, energy policy discussions have been dominated in the past by supply interests. There had been little interest in looking towards more effective energy use as a "source" of energy, even though reinsulation of an attic "supplies" energy to another user willing to pay a higher price. If we take a symmetric view of conservation as part of any energy supply picture, however, and understand how great the potential for energy conservation really is, we should be able to shed the fears of "caves and candles" promised by utility company ads a few years back. As Kenneth Boulding once remarked, "Conservation is just thinking before using energy."

REFERENCES

E. R. Berndt, "Aggregate Energy Efficiency and Productivity Measurement," *Annual Rev. Energy*, 1978, *3*, 225–73, Annual Reviews, Inc., Palo Alto.

——— **and D. Wood,** "Technology, Prices, and the Derived Demand for Energy," *Rev Econ. Statist.*, Aug. 1975, *56*, 259–68.

B. Carlsson, "Structure and Technology: An Analysis of U.S. and Swedish Cement Production," Industrins Utrednings Instit., Stockholm 1977.

J. Gibbons et al., "Report of Demand and Conservation Panel," Committee on Nuclear and Alternative Energy Systems(*CONAES*), National Academy of Sciences, 1978; summarized in "U.S. Energy Demand: Some Low Energy Futures," *Science*, Apr. 14, 1978, *200*, 142–52.

J. M. Griffin and P. R. Gregory, "An Intercountry Translog Model of Energy Substitution Responses," *Amer. Econ. Rev.*, Dec. 1976, *66*, 845–47.

T. V. Long II and L. Schipper, "Resource and Energy Substitution," *Energy*, No. 1, 1978, *3*, 63–79.

S. Özatalay, S. Grubaugh, and T. V. Long II, "Energy Substitution and National Energy Policy," *Amer. Econ. Rev. Proc.*, May 1979, *69*, 369–71.

L. Schipper, "Towards More Productive Energy Use," *Annual Rev. Energy*, 1976, *1*, 455–511, Annual Reviews, Inc., Palo Alto.

——— **and J. Darmstadter,** "The Logic of Energy Conservation," *Tech. Rev.*, Jan. 1978, *80*, 41–50.

——— **and A. Lichtenberg,** "Efficient Energy Use and Well Being: The Swedish Experience," *Science*, Dec. 3, 1976, *194*, 1001–013.

——— **et al.,** study on behalf of the President's Council on Environmental Quality, Lawrence Berkeley Labs, publication no 8299, study underway.

New York Times, Dec. 25, 1977.

Energy Substitution and National Energy Policy

By SAVAŞ ÖZATALAY, STEPHEN GRUBAUGH, AND THOMAS VEACH LONG II*

One issue central to energy policy, planning, and analysis is the extent to which other factors can substitute for energy in the economy. William Hogan and Alan Manne, using a simple dynamic model with a two-input constant elasticity of substitution (*CES*) production function at its core, show that if the elasticity of substitution between energy and an aggregate of all other economic factors is in the range 0.3–0.5, economic growth to the year 2010 is only slightly impeded by even dramatic constraints on growth in energy supply. Conversely, an elasticity of 0.1–0.2 implies a significant depressive effect on the economy if shortages of fuels and electricity occur.

A related topic is the signs and magnitudes of elasticities of substitution between individual factors, such as capital K, labor L, materials M, and energy E. Both Ernst Berndt and David Wood, and Edward A. Hudson and Dale W. Jorgenson have argued on the basis of analyses of time-series data for the *U.S.* economy that energy and capital are complements, while energy and labor are substitutes. James Griffin and Paul Gregory have obtained a conflicting result—that energy and capital are substitutes—utilizing a model developed from pooled cross-national time-series data. All three groups employ a transcendental logarithmic specification of production, but different estimation procedures are used. Because a number of governmental initiatives for energy conservation in the industrial sector are meant to accelerate the introduction of new (and presumably energy-conserving) physical capital, the question of energy-capital complementarity or substitution looms large. If capital and energy are actually net complements, then such policies are misguided and even counterproductive. If capital-energy substitution is facile, then the initiatives would be predicted to have the desired effect.

An ancillary issue is what factor substitutability relations might be anticipated in the medium-to-long run. One method of assessing this is to examine substitution processes in those countries, such as West Germany and Japan, where the relative price of energy has been higher, and more modern energy-conserving technologies are already in place. Although the relative prices of certain factors, such as labor in Japan, may be considerably below that projected for the United States, a thoughtful examination of the data from other nations should provide us with insights.

In this study, we employ pooled cross-national time-series data for seven countries (United States, Canada, West Germany, Japan, the Netherlands, Norway, and Sweden) for the years 1963–74. This data base is utilized in estimating a translog cost function, and from this, Allen partial elasticities of substitution and own- and cross-price elasticities of demand.

I. The Translog Model and Estimation Procedures

The translog model adopted and data construction have been previously described by Gideon Fishelson and Long. Here we investigate a translog cost function of the general form $G = G(Y, P_K, P_L, P_E, P_M)$, in which Y is the output of the total manufacturing sector and P_i is the price of the respective factor. The specific form of the translog cost function is a logarithmic Taylor expansion of this general function, terminated in second order:

*Özatalay is at the Center of Management and Applied Economics, Widener College and the Committee on Public Policy Studies, University of Chicago. Grubaugh and Long are with the Committee on Public Policy Studies, University of Chicago.

$$lnG = \delta_o + \delta_1 lnY + \delta_2(lnY)^2 + \sum_{i=1}^{n} \alpha_i lnP_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \beta_{ij}\, lnP_i lnP_j + \sum_{i=1}^{n} \psi_i lnP_i lnY$$

The cost function is assumed to be well-behaved and to exhibit constant returns to scale. In this model involving capital, labor, energy, and materials, we make no separability assumptions, and these remain to be tested. The model reported here incorporates country dummies, permitting different efficiency parameters. Maximum likelihood estimation techniques are employed.

II. Comparisons with Previous Studies

Table 1 compares our estimates of Allen partial elasticities of substitution with those previously reported. The Griffin-Gregory model is restricted to only three factors—capital, labor, and energy—because comparable international data on materials use is difficult to obtain.

Our data yields a capital-labor elasticity of 1.08, in good agreement with the values obtained by both Berndt and Wood and Hudson and Jorgenson. The very low estimate for this parameter by Griffin and Gregory is counter to economic wisdom and leads one to be suspicious of their results. However, we support their conclusion that capital and energy are good substitutes, a finding diametrically opposed to both of the other studies.

What could underlie this disagreement? A principal difference may be that single country time-series data yield a function that reflects only short-run elasticities, and the link between energy consumption and utilized capital may be strong. Elasticities derived from cross-national data however should reflect longer-run phenomena. The variations in both relative factor prices and cost shares are both smaller for the single country time-series data than for the cross-national time-series data, so that the latter estimates reflect an extension over a much larger portion of the cost function space.

All studies show that labor and energy are effective substitutes, and we also find significant labor-materials and capital-materials substitution. The elasticity of substitution between energy and materials is not significantly different from zero.

From our studies we conclude that factors other than materials appear to be ready substitutes for energy. Thus, we are led to an optimistic appraisal of the potential for continued economic growth, even under moderately severe energy supply restrictions Indeed, it appears that substitution is a most effective means of achieving energy conservation objectives, and it has the additional advantage that it can be implemented in a pricing regime. Capital and energy are shown to be good long-run substitutes, so industrial conservation efforts based on the replacement of vintage facilities seem well-advised.

III. International Comparisons

Allen elasticities of substitution are given for the United States, West Germany and Japan in Table 2. The intercountry differences in the elasticities of substitution are remarkably small, except for the σ_{KK}'s and the large absolute values of σ_{LL} for Japan and σ_{EE} for the United States. The latter two observations appear related to the historically low prices paid for labor in Japan and energy in the United States. It is of interest to observe that the price elasticity of demand for energy in the United States is not significantly

TABLE 1—ALLEN PARTIAL ELASTICITIES OF SUBSTITUTION

Model	σ_{KL}	σ_{KM}	σ_{KE}	σ_{LE}	σ_{LM}	σ_{ME}
This Study	1.08	0.85	1.22	1.03	1.00	0.58
Griffin-Gregory	0.06	–	1.07	0.87	–	–
Berndt-Wood	1.01	0.56	−3.22	0.65	0.60	0.75
Hudson-Jorgenson	1.09	0.25	−1.37	2.16	0.45	−0.77

TABLE 2—INTERNATIONAL COMPARISON OF ELASTICITIES OF SUBSTITUTION[a]

	United States	West Germany	Japan
σ_{KK}	−3.34	−1.89	−2.85
	(.06)	(.03)	(.05)
σ_{KL}	1.08	1.06	1.14
	(.07)	(.05)	(.12)
σ_{KM}	.85	.88	.88
	(.03)	(.03)	(.03)
σ_{KE}	1.22	1.15	1.18
	(.10)	(.09)	(.08)
σ_{LL}	−3.19	−3.51	−7.31
	(.22)	(.25)	(.82)
σ_{LM}	1.00	1.00	1.00
	(.09)	(.12)	(.15)
σ_{LE}	1.03	1.04	1.05
	(.30)	(.45)	(.56)
σ_{MM}	−.83	−1.29	−.60
	(.04)	(.07)	(.03)
σ_{ME}	.58	.42	.65
	(.20)	(.27)	(.16)
σ_{EE}	−32.25	−24.60	−25.14
	(7.75)	(3.27)	(3.54)

[a]Numbers in parentheses are estimated standard errors of the elasticities

smaller than in West Germany and Japan and this again indicates that a pricing policy could play a healthy role in achieving energy conservation objectives. In summary, it appears that a shift by *U.S.* industry to those technologies now employed in greater intensity in Japan and West Germany, while increasing overall efficiency, will not significantly change the flexibility with which the manufacturing sector responds to factor-price changes.

REFERENCES

E. R. Berndt and D. O. Wood, "Technology, Prices and the Derived Demand for Energy," *Rev. Econ. Statist.*, Aug. 1975, *56*, 259–68.

G. Fishelson and T. V. Long II, "An International Comparison of Energy and Materials Use in the Iron and Steel Industry," in Joy Dunkerley, ed., *International Comparisons of Energy Consumption*, Washington 1978, 116–50.

J. M. Griffin and P. R. Gregory, "An Intercountry Translog Model of Energy Substitution Responses," *Amer. Econ. Rev.*, Dec. 1976, *66*, 845–47.

W. W. Hogan and A. S. Manne, "Energy-Economy Interactions: The Fable of the Elephant and the Rabbit?," work. paper EMF no. 1.3, Energy Modeling Forum, Stanford Univ. 1977.

E. A. Hudson and D. W. Jorgenson, "U. S. Energy Policy and Economic Growth, 1975–2000," *Bell J. Econ.*, Autumn 1974, *5*, 461–514.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS OF THE NINETY-FIRST ANNUAL MEETING

CHICAGO, ILLINOIS

AUGUST 29–31, 1978

Minutes of the Annual Meeting
Chicago, Illinois
August 30, 1978

The Ninety-First Annual Meeting of the American Economic Association was called to order by President Tjalling Koopmans at 9:58 P.M., August 30, 1978, in the Grand Ballroom of the Conrad Hilton Hotel. The minutes of the meeting of December 29, 1977 were approved as published in the *American Economic Review Proceedings*, May 1978, pages 440–44.

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), the Managing Editor of the *American Economic Review* (George H. Borts), the Managing Editor of the *Journal of Economic Literature* (Mark Perlman), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were available at registration and were also distributed at the meeting itself. (See their reports published in this issue.) In response to a question from the audience about why the Association pays taxes, the Treasurer responded that some of the activities of the Association that generate revenue are judged by the Internal Revenue Service to be unrelated to the purposes of the Association for which it was given tax exemption status. This unrelated business income (advertising, *Job Openings for Economists*, sales of the mailing lists, etc.) is taxable.

The Secretary presented the following resolutions, which were adopted unanimously:

> BE IT RESOLVED that this meeting record a special vote of gratitude to the members of the 1978 Allied Social Science Associations' Convention Committee chaired by Karl A. Scheld and Roby L. Sloan. The Committee members have devoted many hours of hard work to planning, organizing, and conducting this ninety-first annual meeting of the Association.

> BE IT RESOLVED that this meeting commend Robert Solow, President-elect and 1978 Program Chair, for organizing a varied program of high quality and great interest.

There being no old business, the President called for new business. He stated that three resolutions by members had been submitted to the Secretary a month in advance of this meeting as required by the Association's regulations. These resolutions had been distributed before and at the meeting.

The President called for discussion of the resolution submitted by Wendell C. Gordon and Vincent J. Geraci. The resolution read:

> RESOLVED that those who present Memorial Citations at the Annual Meeting "presidential address session" be respectfully requested to limit their comments to ten minutes.

Thomas Johnson moved the adoption of the resolution and Roger Armondi seconded. Johnson stated that the reading of several "memorials" prior to the President's address at the 1977 New York meetings had been boring and irritating to many members. He recognized that such memorials were customary but desired that they be kept brief. The motion failed.

The President called for discussion of the resolution submitted by William P. Wadbrook and Jerolyn R. Lyle. The resolution read:

> WHEREAS economic and political thought and action are not meaningfully distinguishable, and therefore honors paid to economic or political thought or action on behalf of this Association represent an appropriation and potential abuse of the members' political voices, and therefore such honors, to be just or even valid, must have the express approval of a reasonable proportion of the members, and for purposes of substantive external political action, decisions of a one-year, nonrenewable

panel of officers are not representative of the membership's wishes:

RESOLVED that the By-Laws of this Association be amended as appropriate to provide that functions sponsored by this Association whose avowed purpose is to honor any individuals or entities or activities or ideas must be approved in advance by a two-thirds majority of respondents to a poll in writing.

Since no one present at the meeting moved the adoption of the resolution it died on the floor.

The President called for discussion of the resolution submitted by Paul A. Samuelson and Jere R. Behrman. The resolution read:

WHEREAS Georgia has not ratified the Equal Rights Amendment (ERA) to the U.S. Constitution;

WHEREAS a substantial proportion of the membership prefer not to attend meetings in those states who have not ratified ERA;

WHEREAS many of these members are women and the Association has committed itself to improving opportunities for women in the profession; and

WHEREAS the Association should consider the preferences of the membership in selecting meeting sites:

BE IT RESOLVED that the Association cancel its 1979 meetings in Atlanta and reschedule them in a state that has ratified ERA.

Jere Behrman moved the adoption of the resolution and Ann Friedlaender seconded. Carolyn Shaw Bell moved that the resolution be amended to read:

WHEREAS many of the members of the American Economic Association committed itself to improving opportunities for women in the profession; and

WHEREAS the Association should consider the preferences of the membership in selecting meeting sites;

BE IT RESOLVED that the Association cancel its 1979 meetings in Atlanta and reschedule them in another state.

The motion to amend was seconded. Bell stated that the Certificate of Incorporation of the Association prohibits the Association, as such, from taking a partisan stand on a political issue and that her amendment avoids the political issue raised in the original resolution.

Abba Lerner opposed the amendment on grounds that it was a device to make a political action appear to be nonpolitical. Bell responded by referring to the open letter to members by President Koopmans that appeared in the June 1978 issue of the *American Economic Review* (p. 493–96) in which he reported a decision of the Executive Committee to consider members' preferences in selecting sites for meetings.[1] Bell stated the amendment was designed to make the resolution nonpolitical and reflect members' preferences concerning meeting sites.

Allen Early opposed the amendment on grounds that it would require the Association to violate its contract and it would be difficult to find another site which would satisfy the varied preferences of the membership. Janice Madden responded that sites should be selected based on members' preferences regardless of the basis of those preferences and that other sites could easily be found that were preferable to Atlanta. Marianne Ferber stated that although the Association as such cannot take a partisan stand, the members as individuals can. Members are entitled to express their opinions and expect the Executive Committee to take them into account.

Robert Clower stated that he preferred to meet in cities that do not have antipornography laws; it would be extremely difficult to find sites which were satisfactory to all members. He pointed out that the total membership of the Association is consider-

[1]The Secretary infers that the passage referred to is the following one: "After thorough discussion, the Executive Committee by a decisive majority voted to take no action regarding Chicago or Atlanta but to consider how and to what extent members' preferences concerning site selection could be taken into account in the future" (p. 495).

ably larger than the number present at the business meeting. William Vickrey thought that the amendment carried a lack of candor. Passage of the resolution would be a clear intent to engage in an economic boycott. He felt that a large proportion of the membership would consider the passage of Proposition 13 by California voters a disaster, but this does not rule out California as a possible meeting site for the Association. Candor requires that the resolution be considered a political action.

Werner Hasenberg stated if the purpose of the resolution is not to make a political statement but to ask the Executive Committee to move from Atlanta solely on the basis of the preferences of those members attending the business meeting, the Bell amendment should be amended. He moved and Sushila Gidwani seconded that the resolution be amended to read:

> WHEREAS the Association should consider the preferences of the membership in selecting meeting sites:
>
> BE IT RESOLVED that the Association cancel its 1979 meetings in Atlanta and reschedule them in another state.

Johnson objected to the amendment to the amendment because it was an attempt to camouflage further the political nature of the resolution; adoption of the amendments would be an exercise in self-deception and futility. After some additional discussion in which previously cited reasons for and against the resolution were reiterated, the previous question was moved, seconded, and PASSED by the required two-thirds majority.

In response to a "point of order" concerning correct parliamentary procedure for dealing with amendments to amendments, Fels (the unofficial parliamentarian) explained that a total of three votes on the resolution would have to be taken. The first vote would be whether or not to accept Hasenberg's amendment to Bell's amendment. Regardless of the disposition of Hasenberg's amendment, the house would then need to vote on Bell's amendment. If the Hasenberg amendment passed, the vote on Bell's amendment would be whether to accept or reject it as amended. If the Hasenberg amendment failed, the vote on Bell's amendment would be whether to accept or reject it as originally proposed. Regardless of the disposition of Bell's amendment (amended or not by Hasenberg), the house would have to vote on the original resolution either in its amended form or as originally proposed. When put to a vote, the amendment to the amendment PASSED.

Bell spoke in favor of her amended amendment. She stated that the amendment removed the political implication of the original resolution and that it had been offered in candor for just that purpose. She was not attempting to disguise a political action but to improve the style and wording of the original motion.

Someone asked if the resolution would be submitted to the full membership of the Association if it passed. The President responded that the bylaws empower the Executive Committee to submit to the full membership a resolution adopted at an annual meeting in which less than 5 percent of the membership has participated. The Executive Committee had not made a decision prior to the business meeting.

The previous question was moved, seconded, and passed. The amended motion to amend PASSED.

The original resolution as twice amended was then on the floor for discussion. There being a general state of confusion concerning precisely what question was before the house, the Secretary read the resolution as amended:

> WHEREAS the Association should consider the preferences of the membership in selecting meeting sites:
>
> BE IT RESOLVED that the Association cancel its 1979 meetings in Atlanta and reschedule them in another state.

Madden spoke in favor of the amended resolution. It was clear that the Association was moving solely on the basis of the preferences of its members. If not going to Atlanta was interpreted as taking a political position, going could also be so interpreted. The Asso-

ciation was making a political statement regardless of what action it took.

At this point, President Koopmans briefly reviewed the history of the issue as discussed in the March 17, 1978 and December 27, 1977 meetings of the Executive Committee, summarized his open letter to members which appeared in the June 1978 issue of the *AER*, and read exerpts from that letter. (For the minutes of the December 27, 1977 meeting, see pages 449–52 of the May 1978 *AER*; for March 17, 1978, see this issue; and for the President's letter, see the June 1978 issue.) He then asked Counsel Leo Raskind for advice concerning whether or not the amended resolution was in order. Counsel advised that in his opinion the resolution was in conflict with article three of the Association's Certificate of Incorporation. The legal effect of passage of the resolution would be to breach a contract; damages could be substantial Because the courts would consider not just the resolution but also the surrounding discussion, minutes of the meetings of the Executive Committee, and testimony as relevant evidence, it seemed clear to him that the resolution was in violation of the Association's Charter. The President did not rule the resolution out of order.

Vickrey stated that wholly aside from the Association's attempt to refrain from taking political stands, passage of the resolution would be an attempt to subvert the political process of Georgia. For us to boycott by threatening economic loss unless citizens of Georgia adhere to our preferences, would be in effect attempting to bribe them.

Ann Friedlaender said that the motivation behind the resolution was simply to correct an error that had been made when Atlanta was selected in 1973. Members prefer to meet somewhere other than Atlanta. Clower responded that it was not a mistake. Given an extension of time allowed for passage of the Equal Rights Amendment, Georgia and Illinois may still ratify the amendment. Therefore the resolution is empty and makes the Association look foolish.

The President asked the Secretary to outline the problems that might arise if the Association voted to move the meetings from Atlanta. The Secretary responded that the first problem would be finding another appropriate site. He understood the spirit of the resolution to preclude such cities as Chicago, St. Louis, Kansas City, and New Orleans. Consequently, the meetings would probably have to be held on either the East or West Coast. Given the lead time necessary to investigate convention sites and negotiate contracts, it may be that other Associations have already booked hotel space in the cities most appropriate for AEA meetings. He did not know if space would be available.

The second problem would be arranging for local support to help organize and manage the meetings. The Association relies heavily on such support. The Atlanta Convention Committee has already been formed, and many of its members are present at these meetings learning their responsibilities. The Secretary despaired of finding another committee on such short notice. He believed that additional people would have to be hired in his office in order to handle the job.

The third problem was of a long-run nature. Hotels, confronting higher risks, would adjust prices accordingly. New contracts would probably be less favorable in terms of room rates, costs of public space, and other hotel and convention services.

Robert Solow asked Behrman and Bell if they would be willing to change the resolution to indicate that the Association regretted the Executive Committee's decision to hold the meetings in Atlanta. The response was negative.

Moses Abramovitz pointed out that a large part of the program consists of sessions organized by the allied associations. Unless the AEA's action was accompanied by similar actions on the part of the allied associations, organizing a program would be more difficult than usual—more difficult, not impossible. In addition, he distinguished between choosing among cities for future meetings and a cancellation. Preferences certainly have a place in selecting future sites and accounting for such preferences would not be perceived as a political action. The only interpretation of the resolution before the house was that the Association was sufficiently in sympathy with

some of its members' political preferences to support them.

William Hellmuth stated that the Association was objecting to a state of affairs that it could not change. Smaller attendance at the annual meetings is a lesser evil than breaching a contract. The decent and prudent thing to do is for the Association to go to Atlanta, honor its commitment, but leave to each member the decision of attending or not.

Bruce Hamilton moved the previous question, and the motion PASSED. The resolution as twice amended failed: 56 for, 57 against.[2]

It was moved that the Association not schedule any future meetings in states that have not ratified the equal rights amendment. The Chair ruled the motion out of order on grounds that it violated the articles of incorporation of the Association.

The Chair then introduced Robert Solow, the President-elect. There being no further business, the meeting adjourned at 12:25 P.M.

C. ELTON HINSHAW, *Secretary*

[2]The 57 "No" votes included that of the President. Members were asked to indicate their vote by standing. The Secretary and Treasurer, acting as tellers, went to the floor to count. As they returned to the podium, the President asked if they had counted his vote which was "No." The tellers indicated they had not. Since the vote had been a tie (56 to 56), the President's vote was counted. The Secretary announced the final count as 56 for, 57 against including the President's vote. A check on parliamentary rules after the meeting revealed that the motion should have been deemed to have lost without the President's (negative) vote: "On a tie vote, a motion requiring a majority vote for adoption is lost, since a tie is not a majority. Thus, if there is a tie without the chair's vote, the presiding officer can, if he is a member, vote in the affirmative, thereby causing the motion to be adopted, or, if there is one more in the affirmative than in the negative without the chair's vote (for example, if there are 72 votes in favor and 71 opposed), he can vote in the negative to create a tie, thus causing the motion to be rejected."

Minutes of the Executive Committee Meeting

Minutes of the Meeting of the Executive Committee in New York, New York, March 17, 1978.

The first meeting of the 1978 Executive Committee was called to order at 9:25 A.M. on March 17, 1978 in the New York Hilton Hotel, New York, New York. The following members were present: Tjalling Koopmans (presiding), George H. Borts, Robert W. Clower, Edward Denison, Rendigs Fels, C. Elton Hinshaw, Robert J. Lampman, Franco Modigliani, Marc Nerlove, Joseph Pechman, Mark Perlman, Edmund Phelps, Alice Rivlin, Robert Solow, and Marina v.N. Whitman. Present as counsel was Leo J. Raskind. Present as members of the Nominating Committee for a part of the meeting were Andrew F. Brimmer, Carolyn Shaw Bell, Robert Haveman, Bert G. Hickman, and Vernon Smith. Present as guests for parts of the meeting were Susan Rose-Ackerman, Marcus Alexis, Ann Friedlaender, Michael Lovell, Edwin S. Mills, Wyn Owen, Lloyd G. Reynolds, and Wilma St. John.

Koopmans called the meeting to order, reviewed the agenda, and noted additions to it.

Minutes. The Secretary made some minor corrections in the minutes of the meeting of December 27, 1977 which had been previously circulated. The corrected minutes were approved.

Report of the Secretary (Hinshaw). An oral report was not given. The Secretary's written report had been previously circulated. The report stated that the 1978 meetings would be held at the Conrad Hilton in Chicago on August 29–31. The schedule for subsequent meetings is: December 28–30, 1979 in Atlanta; September 5–7, 1980 in Denver; and December 28–30, 1981 in Washington, D.C. The site for 1982 will be New York, but the dates have not yet been selected. Tentative plans are to hold the 1983 meetings in San Francisco on December 28–30. The report also stated that the Secretary had assigned the duties of convention management to Barbara Weaver, Assistant Administrative Director in the Nashville office.

The Secretary has entered into a contract with R.R. Donnelley & Sons to print the 1978 handbook of members. The Ad Hoc Committee (James Morgan, Joseph Pechman, Barbara Reagan, and Hinshaw) appointed to design the questionnaire had approved the one to be used and a copy was contained in the report. The questionnaire will be mailed to the members in late March or early April. The handbook is scheduled for publication and mailing in November or December. The Secretary hopes to begin publishing a handbook every three years.

Report of the Editor of the American Economic Review (Borts). Borts reported on the editing of the *Papers and Proceedings*. On his recommendation, the following persons were elected to the Board of Editors of the *American Economic Review* for three-year terms: Rudiger Dornbusch, William H. Oakland, Richard W. Roll, A. Michael Spence, William S. Vickrey, and S. Y. Wu.

Report of the Editor of the Journal of Economic Literature (Perlman). Perlman reported that the 1974 *Index of Economic Articles* would be published this summer and the 1975 and 1976 volumes would be published during 1979, that a contract proposal had been received from Lockheed Missiles & Space Company, Inc. to provide online information retrieval activities to third parties, and that beginning with the March issue, *JEL* would contain a list of authors of journal articles along with the subject categories of their articles.

Report of the Treasurer (Fels). The Treasurer reported that the surplus for 1977 was $156 thousand and the net worth of the Association is now $431 thousand, about 43 percent of the 1978 budgeted expenditures. It appears that by the end of 1978 the net worth will approximate the cumulative deficits of 1969–75 and therefore may be considered an adequate reserve. But in the normal course of events the surplus will decline after 1978 at a rate of $50 thousand a year, perhaps more. At

that rate there will be a deficit in 1981 unless dues and subscription rates are raised.

Committee on the Status of Minority Groups in the Economics Profession (Alexis). Alexis reported that the 1978 summer program for minority students is in jeopardy of being discontinued unless the Sloan Foundation grant is matched by other funds. The Foundation has awarded $170 thousand for the summer programs of 1977–79 on the condition that matching funds are raised. The Foundation agreed to release the first $55 thousand on the condition that an equal amount be raised before the next allocation of $55 thousand would be released. To date $30 thousand has been raised ($20 thousand of AEA funds and $10 thousand of Ford Foundation funds). An additional $25 thousand is needed in order to have the Sloan Foundation release the $55 thousand for the 1978 summer program. The remaining funds would be used for the 1979 program. After a discussion of the long-run financial problems and the value of the program, it was VOTED to appoint a committee to evaluate the existing effort of the Association in recruiting minorities into the profession and to examine alternative approaches.

Economics Institute (Owen and Mills). Owen reported that the Economics Institute is following a tuition and fee policy that aims to obtain full reimbursement from students with well-endowed sponsors, and the Institute is now about 90 percent self-supported. However, in pursuing this policy, the need for fellowship funds to help finance applicants with less sponsor support has increased. The Institute could effectively use about $25 thousand each year, and it is seeking fellowship funds from several potential sources in the interest of such applicants. It was VOTED to appropriate a terminal grant of $5 thousand in fellowship funds for the Institute with the proviso that it would continue to seek an optimal mix of students from various countries.

Search Committee for Editors (Lovell). In the absence of James Tobin, the chairman, Lovell presented the report of the Committee. In addition to searching for new editors of the *American Economic Review* and the *Journal of Economic Literature*, the Committee has discussed the editorial management structure of the two journals. In the case of the *JEL*, the Committee had previously recommended and the Executive Committee had approved the continuation of Naomi Perlman as Associate Editor for an additional four-year term. She will manage the journal indexing, the abstracting of articles, and the publication of the annual indexes. For the *AER*, the Committee recommended the hiring of three co-editors in addition to the Managing Editor; not more than one from the same institution as the Managing Editor. The nature of these positions and the persons to be appointed would be worked out with the new editor. The Executive Committee agreed but took no formal action.

Lovell submitted a rank ordered list of nominees for the editorships—three for the *AER* and four for the *JEL*—and stated that the Committee deliberately had not contacted any of the persons recommended to ascertain the extent of their interest or their availability. It was VOTED to ask the Committee to explore the availability of the nominees for editorship of the *AER* and give another report on the criteria used to establish their rank ordering of the nominees. It was understood that the exploration would be done simultaneously with the three nominees. It was VOTED to approve the first two names (in rank order) submitted for the editorship of the *JEL*. If neither accepts, the Committee was instructed to submit a new slate of nominees to the Executive Committee.

Committee on U.S.–Soviet Exchange (Reynolds). Reynolds reported that an exchange visit to the Soviet Union had been arranged for June 24–July 2, 1978, and that the visit will be funded by the State Department. The topic of the symposium is "Problems of Industrial Management." The Committee continues to be concerned with increasing the breadth and depth of research cooperation.

The Equal Rights Amendment (ERA) and Site Selection (Friedlaender and Rose-Ackerman). Before introducing Friedlaender and Rose-Ackerman, Koopmans explained that the question was being reconsidered

because of the concerns of a significant number of members expressed to him since the last meeting of the Executive Committee. Rose-Ackerman and Friedlaender proposed that the Executive Committee (1) explore the possibility of moving the Atlanta meeting to a state that has ratified ERA and (2) issue a statement to the effect that after 1979 the Association will not meet in states that have not ratified the ERA so long as passage of the Amendment is a live issue.

After their presentation, the Executive Committee discussed the subject at length. As far as could be told from the discussion, there was full agreement on two points; it was too late to change the site of the meetings scheduled for August 29–31, 1978, in Chicago; and the provision in the charter forbidding the Association from taking a position on partisan issues was wise. Those favoring a move out of Atlanta and a policy of refusing to meet in states that have not ratified ERA expressed the view that the commitment to meet in Atlanta may not be a firm contractual obligation, that the likelihood of the Atlanta hotels bringing a lawsuit under the laws of contract or the antitrust laws was minimal, that the Association could take into account the preferences of its members as to meeting sites without violating its charter (i.e., without taking a position on ERA itself), that meeting in Atlanta might discourage attendance by young women economists most in need of the benefits of attending or put them in the awkward position of compromising their principles, that meeting in Atlanta would be inconsistent with the position taken by the Association and has been ratified by states with 70 percent of the population of the United States, that the Association can most effectively make known the position of its members on ERA by refusing to meet in states that have not ratified it, and that in view of the number of associations that have taken stands of the kind being proposed for the AEA, failure to act in this way would be construed as a negative position on ERA itself. Opposition to the proposals centered on three arguments: that it is illegitimate for the Association to use economic pressure to bring about political change, that adoption of the proposals would be taken as endorsement by the AEA of ERA, i.e., would constitute taking a position on a political issue, and that selecting sites to further a political measure would set a dangerous precedent. It appeared from the discussion, however, that there was considerable agreement that the Association in future selection of meeting sites should take account of the preferences of its members, regardless of their basis.

After a thorough discussion, it was VOTED to take no action regarding Chicago or Atlanta but to consider how and to what extent members' preferences concerning site selection could be taken into account in the future.

Nominating Committee (Brimmer). The Electoral College consisting of the Nominating and Executive Committees meeting together chose Moses Abramovitz as the nominee for President-elect and elected Richard Musgrave and William Vickrey Distinguished Fellows. On behalf of the Nominating Committee, Brimmer reported the following nominees for other offices in the 1978 election: for Vice President (two to be elected), Hollis Chenery, Arnold Harberger, Jack Hirshleifer, and Irma Adelman; for members of the Executive Committee (two to be elected), Henry Aaron, Samuel Bowles, Zvi Griliches, and Daniel McFadden.

Ad Hoc Committee on "Successions" (Lampman). Lampman reported that after carefully reviewing the existing Association procedures for responding to vacancy in the office of President-elect and many alternatives, the Committee concluded that the existing bylaws and procedures are adequate. The Committee offered the following motion regarding the succession to a vacancy in the office of President-elect:

> The Executive Committee hereby resolves that it is presumed that in the case of death or disability of the President-elect the successor to the President-elect of this Association will be the person previously named as alternate for the office of President-elect by the most recent electoral college.

It was VOTED to accept with approval the Committee's report. No further action was

taken on the report. However, it was VOTED that it is presumed that in case of the death or disability of the nominee for President-elect the Executive Committee shall consider the "alternate" as the new nominee and submit the candidate to the membership for election.

1978 Program (Solow). The President-elect reported that the program was almost complete. He sought advice about cosponsoring sessions with groups that are not officially affiliated with the Allied Social Science Associations. He was advised that caution should be taken in cosponsoring such sessions.

Federal Funding of Research (Koopmans). The President announced that Lampman had agreed to chair the Committee on Federal Funding of Research. It was understood that this was not a "lobbying" Committee, but that testifying before Congress would be a legitimate function of the Committee.

The meeting adjourned.

Minutes of the Meeting of the Executive Committee in Chicago, Illinois, August 28, 1978.

The second meeting of the 1978 Executive Committee was called to order at 10:15 A.M on August 28, 1978 in the Conrad Hilton Hotel, Chicago, Illinois. The following members were present: Tjalling Koopmans, (presiding), George H. Borts, Robert W. Clower, Edward Denison, Rendigs Fels, C. Elton Hinshaw, Robert J. Lampman, Marc Nerlove, Joseph Pechman, Mark Perlman, Edmund Phelps, Robert Solow, and Marina v.N. Whitman. Also present were Leo J. Raskind, counsel and Moses Abramovitz, nominee for President-elect. Present as guests for parts of the meeting were Marcus Alexis, Robert Ferber, Dwight Perkins, Lloyd Reynolds, Carl Stevens, and James Tobin.

Minutes. The minutes of the March 17, 1978 meeting were approved as written and circulated.

Report of the Secretary (Hinshaw). The Secretary reported that the 1979 annual meetings are scheduled to be held in Atlanta, Georgia on December 28–30 with headquarters at the Atlanta Hilton Hotel. The schedule for subsequent meetings is: September 5–7, 1980 in Denver, December 28–30, 1981 in Washington, D.C., and 1982 in New York. Hotel space is currently reserved for both the Labor Day period and the Christmas–New Year period. San Francisco has tentatively been selected as the site for 1983.

He also reported that, effective May 29, 1978, second class postage costs increased. Estimated postage for 1978 is $4.25 per foreign member and subscriber for the journals and $.85 per foreign member for the 1978 *Handbook*. The Association currently charges $3.70 for postage; 1978 costs will not be recovered even excluding the *Handbook*. Estimated foreign postage for 1979 is $4.45 per foreign member and subscriber. The Executive Committee accepted his recommendation to raise the foreign postage charge to $4.75 for 1979. This rate will cover the costs of mailing the journals and a prorated share of the costs of mailing the *Handbook*.

Report of the Treasurer (Fels). The Treasurer reported that the Association had a surplus of $156,000 in 1977, but he projected a decline of $50,000 per year, implying a need to raise dues effective January 1, 1981. Results for the first half of 1978 indicate that his projection is on target. The recommended 1979 budget will be circulated to the Executive Committee by mail. Full discussion of the budget should take place at the spring 1979 meeting.

The Executive Committee approved the Treasurer's recommendation to move the bank account of the Association from the State National Bank of Evanston to Commerce Union Bank of Nashville and to move the custodian account from the State National Bank of Evanston to the La Salle National Bank of Chicago.

Report of the Editor of the American Economic Review (Borts). The Editor reported that submissions continue to decline. They reached a peak in 1970, and on average have been declining each year since. When questioned about the effect of page and submission charges on the number of submissions, Borts responded that he thought the decline was due to the appearance of new journals and the review time necessary for publication decisions rather than page and submission charges.

Report of the Editor of the Journal of

Economic Literature (Perlman). On recommendation of the Editor, Edwin Burmeister and Roy Weintraub were elected to the Board of Editors. Perlman stated that he was leaving two places on the Board vacant so that the new editor could recommend their appointment. He is working on the assumption that he will not be the editor in 1980 and has planned only four more issues so that the next editor will not have commitments outstanding. He announced the publication of the 1974 *Index of Economic Articles* and indicated that the 1975 and 1976 volumes would appear next year. He expressed disappointment with the sales volume of the *Indexes*, particularly since its price was lower than similar volumes in the field of political science.

Report of the Director of Job Openings for Economists (Hinshaw). The Director reported that the number of vacancies listed in the first four issues in 1978 was essentially the same as last year, the division between academic and nonacademic jobs continued to be about two to one, and general economic theory was the field of specialization most in demand.

Search Committee for Editors (Tobin). Acting on the Search Committee's recommendation, the Executive Committee voted to offer the position of Managing Editor of the *American Economic Review* to its first candidate. For Managing Editor of the *Journal of Economic Literature*, the Committee suggested several possibilities. After a discussion of the strengths and weaknesses of each of the possible candidates, the Executive Committee voted to approve a rank-ordered slate of candidates and authorized the President, in concert with the Chairman of the Search Committee and the President-elect, to offer the Managing Editorship to persons on the list in the order of their ranking.

Committee on Publications (Ferber). Ferber reported on the recent history of the Committee and sought advice and instruction as to its proper function. It was agreed that the main function is to oversee the journals. It was also suggested that the Committee should investigate the desirability of translations and a series similar to Cambridge University Press' on Keynes and Jevons (for example, Irving Fisher).

Committee on U.S.–Soviet Exchanges (Reynolds). Reynolds reviewed the Symposuim held in the *USSR* in June 1978. Negotiations are currently underway for a return visit to the United States. He reported that funding is still uncertain, three different sources have supported the last symposia, and long-term financing of the exchanges needs to be obtained.

Committee on U.S.–China Exchanges (Perkins). Perkins reported that for a variety of reasons this is a propitious time to attempt to establish an exchange program with China. It was voted to initiate an exploration of the possibility of establishing exchanges with China and to continue the Soviet exchange program at the same level. It was understood that the present Committee would be expanded to make it more representative of the Association.

Ida Nudel (Klein). Kenneth Arrow had written a letter requesting the Executive Committee pass a resolution to the following effect:

> The American Economic Association regrets the arrest and conviction of the economist, Ida Nudel, on grounds that clearly represent solely her expression of free speech and freedom of movement and in no way represent a threat to a legitimate government. In the interests of intellectual cooperation among the countries of the world, we urge the government of the Soviet Union to commute Ida Nudel's sentence.

He asked that the resolution be sent to Anatoly F. Dobrynin, Ambassador Extra-Ordinary of the *USSR* and Leonid Brezhnev, President of the *USSR*. An amended version of Arrow's resolution was moved and seconded. To wit:

> The Executive Committee of the American Economic Association regrets the arrest and conviction of the economist, Ida Nudel, on grounds that clearly represent solely her expression of free speech and freedom of movement and

> represent a denial of her human rights. In the interests of intellectual cooperation among the countries of the world, we urge the government of the Soviet Union to commute Ida Nudel's sentence.

After an extended discussion of the appropriateness of the Executive Committee taking stands on rather randomly discovered individual cases such as this, the relationship of the resolution to the Association's charter which prohibits partisan attitudes, and the probable effect of the resolution, it was passed (five for, three against, two abstentions). Whereupon it was voted unanimously to reconsider. The original motion was then withdrawn. It was agreed that individual members of the Executive Committee would sign a joint letter concerning Nudel.

Committee on Political Discrimination (Stevens). The Chairman reported that the main business of the Committee has been to investigate individual complaints. Such complaints have diminished in the last few years. Little progress has been made in advancing the research proposal. The project would cost more than the $10,000 voted by the Executive Committee at its December 27, 1977 meeting. Stevens asked if the Executive Committee's allocation of $10,000 meant a lack of enthusiasm for the project. He was informed that the $10,000 was the amount allocated to the Committee on the Status of Women in the Economics Profession and the Committee on the Status of Minorities. The Executive Committee was seriously concerned about the issue but had doubts about how fruitful such research would be.

Committee on the Status of Minority Groups in the Economics Profession (Alexis). Alexis reported on the fifth summer program; 25 students came; 24 completed; and 23 completed satisfactorily. Twenty-two schools were represented. The budget was approximately $77,000 of which $55,000 came from the Sloan Foundation, $20,000 from the Association, and the remainder from a university that supported its student. He was not requesting any additional funds from the Association for the 1978–79 program. Acting on his recommendation, the Executive Committee voted to endorse the concept of a consortium (consisting of Yale, M.I.T., Northwestern, Stanford, and possibly Princeton and the University of California-Berkeley) to support the training of minority students.

Pro Forma Publications Contract (Koopmans). It was voted to establish a committee to survey the current practices of publishers in contracting for manuscripts with members of the Association.

Ad Hoc Committee on Federal Funding of Economic Research (Klein). Klein raised the matter of the Association taking a more vigorous approach to securing federal funding of basic research in economics and the social sciences. Lampman stated that the existing Committee on Federal Funding would have a full report to the Executive Committee in about a year.

Resolutions from Members (Koopmans). The resolutions submitted by members for consideration at the annual business meeting were reviewed and discussed.

The meeting was adjourned at 11:35 P M

C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1978

Annual Meetings. In 1979 the annual meetings will be held at the Atlanta Hilton Hotel in Atlanta, Georgia on December 28–30. The schedule for subsequent meetings is: September 5–7, 1980, in Denver, Colorado; December 28–30, 1981, in Washington, D.C.; and December 28–30, 1982, in New York, New York. The Executive Committee has made a tenative decision to meet in San Francisco, California, December 28–30, 1983.

Employment Services. For those meetings scheduled for December 28–30, employment services will be provided at the annual meeting but will begin December 27. Because the 1980 annual meetings will occur at an early period in the academic year, it was decided to provide employment services at a later time. The dates and site of the 1980 placement meeting have not yet been selected. Formal placement services will not be provided at the September 1980 meetings.

The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. All economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision of the employment service porvided at the annual meetings.

Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and of their professional obligation to list their openings.

Membership. The total number of members and subscribers, shown in Table 1, reached an all time high of 26,787 at the end of 1975. The introduction of a progressive dues structure in 1976 may account for most of the decline in the number of members and subscribers in 1976 and 1977. The desire to be listed in the 1978 Directory may account for the increase in members this year.

Permission to Reprint and Translate. Official permissions to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 303 in 1978 compared to 236 in 1977. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

Visiting Economics Scholars Program. The purpose of the Visiting Economics Scholars Program is to facilitate visits by leading economists to smaller colleges emphasizing teaching. The host colleges are expected to pay part or all of the costs of the visits; at a minimum they take care of the local expenses and travel costs of the visitors. During the academic year 1977–78 there were two such visits sponsored by the program.

Directory. A special issue of the *American Economic Review* was published and distributed to members late this year. It contained articles and material that has usually been published separately as a directory or handbook. Postal regulations prohibit the mailing

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of Year)

	1976	1977	1978
Class of Membership			
Annual	15,102	14,379	15,698
Junior	2,631	1,731	1,857
Life	399	375	389
Honorary	36	33	35
Family	344	284	307
Complementary	560	537	615
Total Members	19,072	17,339	18,901
Subscribers	7,134	6,728	6,893
Total Members and Subscribers	26,206	24,067	25,794

of freestanding directories at second class postal rates. Special issues of the *Review* qualify for second class mailing if they include material that appears in the regular issues. The decision to change the format of the "Directory" issue of the journal decreased the total costs of publishing and distributing the Directory.

Committees and Representatives. Listed below are those who served the Association during 1978 as members of Committees or representatives. The year in parenthesis indicates the final year of the term to which they have been appointed most recently. On behalf of the Association, I wish to thank them all for their services.

AD HOC COMMITTEE TO DESIGN A QUESTIONNAIRE FOR THE 1978 HANDBOOK
James N. Morgan
Joseph Pechman
Barbara Reagan
C. Elton Hinshaw, *ex officio*

AD HOC COMMITTEE TO EVALUATE THE ASSOCIATION'S PROGRAM IN ATTRACTING MINORITIES INTO THE PROFESSION
Albert Rees, *Chair*
Bernard E. Anderson
John U. Monro
Finis Welch

AD HOC COMMITTEE TO INVESTIGATE THE FEASIBILITY AND DESIRABILITY OF ESTABLISHING AN EXCHANGE PROGRAM WITH CHINA
Dwight Perkins, *Chair*
John G. Gurley
Benjamin Ward

AD HOC ADVISORY COMMITTEE TO THE NATIONAL COMMISSION ON EMPLOYMENT AND UNEMPLOYMENT STATISTICS
Harold Watts, *Chair*
Orley Ashenfelter
Carolyn Shaw Bell
Charles C. Holt

AD HOC COMMITTEE ON PUBLISHING CONTRACTS
Martin Shubik, *Chair*
William J. Baumol
Leo Raskind
C. Elton Hinshaw, *ex officio*

AD HOC COMMITTEE TO REVIEW NEW PROPOSED STANDARD OCCUPATIONAL CLASSIFICATION SYSTEM
H. Gregg Lewis, *Chair*
Victor Fuchs
Margaret S. Gordon
Michael Piori
Sherwin Rosen

AD HOC COMMITTEE TO REVIEW PROCEDURES FOR REPLACEMENT OF PRESIDENT AND PRESIDENT-ELECT
Robert J. Lampman
Burton Weisbrod

BUDGET COMMITTEE
Edmund S. Phelps, *Chair* (1978)
Marina v.N. Whitman (1980)
Robert J. Lampman (1979)
Rendigs Fels, *ex officio*
Tjalling C. Koopmans, *ex officio*
Robert Solow, *ex officio*

CENSUS ADVISORY COMMITTEE
James R. Nelson, *Chair* (1979)
Nancy S. Barrett (1980)
Victor R. Fuchs (1980)
George L. Perry (1980)
Norman Simler (1980)
Lester C. Thurow (1980)
Carolyn Shaw Bell (1979)
Andrew F. Brimmer (1979)
Burton Malkiel (1979)
Arnold Zellner (1979)
Barbara Bergmann (1978)
Anne P. Carter (1978)
Robert F. Lanzillotti (1978)
William Niskanen (1978)
Richard Ruggles (1978)

COMMITTEE ON ECONOMIC EDUCATION
Allen C. Kelley, *Chair* (1979)
George Leland Bach (1980)
William E. Becker (1980)
Keith Lumsden (1980)
Walter Heller (1979)
Elizabeth Allison (1978)
W. Lee Hansen, *Acting Chair* (1978)
John Siegfried (1978)
Rendigs Fels, *ex officio*

ECONOMICS INSTITUTE POLICY AND ADVISORY BOARD
Edwin S. Mills, *Chair* (1981)
John Day (1981)
Axel Leijonhufvud (1981)
Carlos F. Diaz-Alejandro (1980)
Raymond Vernon (1980)
Carl Keith Eicher (1979)
Anne O. Krueger (1979)
Paul G. Clark (1978)

COMMITTEEE ON ELECTIONS (1978)
Ben Bolch, *Chair*
Barbara Haskew
C. Elton Hinshaw, *ex officio*

COMMITTEE ON FEDERAL FUNDING OF ECONOMIC RESEARCH
Robert J. Lampman, *Chair* (1979)
Richard Freeman (1979)
Anne O. Krueger (1979)
George L. Perry (1979)
Richard N. Rosett (1979)

FINANCE COMMITTEE
Robert Eisner, *Chair* (1980)
Robert G. Dederick (1979)
James Lorie (1978)
Rendigs Fels, *ex officio*

COMMITTEE ON HONORARY MEMBERS
Leonid Hurwicz, *Chair* (1980)
William J. Baumol (1982)
Hollis B. Chenery (1982)
Paul A. Samuelson (1980)
Bent Hansen (1978)
W. Arthur Lewis (1978)

COMMITTEE ON HONORS AND AWARDS
Irma Adelman, *Chair* (1978)
Moses Abramovitz (1982)
Carl F. Christ (1982)
John Chipman (1980)
James W. McKie (1980)
Marcus Alexis, *Acting Chair* (1978)

NOMINATING COMMITTEE (1978)
Andrew F. Brimmer, *Chair*
Carolyn Shaw Bell
Peter A. Diamond
John G. Gurley
Robert M. Haveman
Bert G. Hickman
Vernon L. Smith

COMMITTEE ON POLITICAL DISCRIMINATION
Carl M. Stevens, *Chair* (1980)
Anne P. Carter (1981)
Kenneth J. Arrow (1980)
John G. Gurley (1980)
Robert E. Lucas, Jr. (1980)
William J. Baumol (1978)

COMMITTEE ON PUBLICATIONS
Robert Ferber, *Chair* (1978)
Edwin Burmeister (1979)
Peter A. Diamond (1979)
Robert Gallman (1978)
John G. Gurley (1978)
Michael Lovell (1978)
C. Elton Hinshaw, *ex officio*

SEARCH COMMITTEE FOR EDITORS
James Tobin, *Chair*
Irma Adelman
Albert Ando
George R. Feiwel
Martin S. Feldstein
John G. Gurley
Michael C. Lovell
Bernard Saffran

COMMITTEE ON THE STATUS OF MINORITY GROUPS IN THE ECONOMICS PROFESSION
Marcus Alexis, *Chair* (1980)
James N. Morgan (1981)
Guy H. Orcutt (1980)
Albert Rees (1980)
Andrew F. Brimmer (1979)
Alice Rivlin (1979)

COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION
Ann Friedlaender, *Chair* (1979)
Marianne Ferber (1980)
Ruth Gilbert Shaeffer (1980)
Miriam K. Chamberlin, *Vice-Chair* (1979)
William F. Hellmuth (1979)
Janice Madden (1978)
Margaret C. Simms (1978)
Tjalling C. Koopmans, *ex officio* (1978)

COMMITTEE ON U.S.–SOVIET EXCHANGES
Lloyd G. Reynolds, *Chair* (1979)
Abram Bergson (1979)
John Meyer (1979)
Rendigs Fels, *ex officio*

Council and Other Representatives

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
William Nordhaus (1979) Section on Social and Economic Sciences

AMERICAN COUNCIL OF LEARNED SOCIETIES
William Parker (1978)

FEDERAL STATISTICS USERS CONFERENCE
Edward F. Denison (1979)

INTERNATIONAL ECONOMIC ASSOCIATION
Abram Bergson (1978)
Franco Modigliani (1981)

INTERSOCIETY COMMITTEE ON TRANSPORTATION
William Dodge

POLICY BOARD OF THE JOURNAL OF CONSUMER RESEARCH
Kelvin J. Lancaster (1978)

NATIONAL ARCHIVES ADVISORY COUNCIL-GENERAL SERVICES ADMINISTRATION
Robert Gallman (1978)

NATIONAL BUREAU OF ECONOMIC RESEARCH
Carl F. Christ (1981)

SOCIAL SCIENCE RESEARCH COUNCIL
Robert Eisner (1978)

Representatives of the Association on Various Occasions–1978

INAUGURATIONS

John W. White, Jr., Nebraska Wesleyan University
Theodore W. Roesler

Arthur H. DeRosier, Jr., East Tennessee State University
Eugene P. Price

Joseph M. McFadden, Northern State College
Frank W. Smith

James Thomas Laney, Emory University
James F. Crawford

James Thomas McComas, Mississippi State University
Michael P. O'Neill

Rutherford H. Adkins, Knoxville College
Gordon W. Ludolf

Charles E. Glassick, Gettysburg College
Roswell G. Townsend

Cleveland L. Dennard, Atlanta University
Barbara A. Jones

Charles D. Lein, The University of South Dakota
Kenneth L. Bauge

Stephen J. Trachtenberg, The University of Hartford
Richard Scheuch

Michael P. Hammond, The State University of New York–College at Purchase
Seamus O'Cleireacain

Thomas N. Bonner, Wayne State University
Donald R. Byrne

William S. Banowsky, The University of Oklahoma
Upton B. Henderson

Elias Blake, Jr., Clark College
James F. Crawford

Matthew E. Creighton, Creighton University
William R. Hosek

Hanna Holborn Gray, The University of Chicago
John F. McDonald

DIAMOND JUBILEE CELEBRATION OF TEXAS WOMAN'S UNIVERSITY
Kendall P. Cochran

1978 ASSA Convention Committee

Karl A. Scheld, *Chair*
Roby L. Sloan, *Co-Chair*
Edward H. Boss, Jr.
Richard Chamberlin
Robert G. Dederick
Bert E. Elwert
Walter D. Fackler
Nancy M. Goodman
William L. Helfers
Jim Lilly
Laurence Jay Mauer
Morton B. Millenson
Margaret Oppenheimer
Richard S. Peterson
Julian Schwimmer
Violet O. Sikes
George Spink
William R. Waters
Barbara Weaver

1979 ASSA Convention Committee

Harry Brandt, *Chair*
John M. Godfrey, *Vice-Chair*
James C. Armstrong
William E. Black
Kong Chu
James E. Clark
Donald M. Cruse
Arnold A. Dill
M. M. Galloway
Marlin V. Law
W. Bethel Minter
Cindy Russell
Arthur F. Schreiber
Jack R. Sicard
Violet O. Sikes
Barbara Weaver
Philip M. Webster

C. ELTON HINSHAW, *Secretary*

Report of the Treasurer for the Year Ending December 31, 1978

The financial situation of the American Economic Association continues to be strong. Its net worth at the end of 1978 was $633 thousand, about 60 percent of expected expenditures for 1979. (See balance sheet appended to the Auditors' Report and Table 1 of this report.) This is adequate protection against unexpected developments. During the next few years, the net worth will continue to rise as a result of continued surpluses. The desirable size of the net worth will also continue to rise as inflation raises expenditures. These offsetting tendencies will result in the net worth continuing at a comfortable ratio to annual expenditures. In the near term there is no need for dues increases.

The income statement for 1978 shows another large surplus. The apparent rise in the surplus, however, is spurious. In the transition from a hand-operated system to computerization in 1977, the mailing of second notices to members who had not paid their dues got delayed. As a result, an unusually large amount of dues applicable to 1977 was received too late in 1978 for inclusion in the audited financial reports for the previous year and had to be included in the 1978 data. There is no way to tell the exact amount (nor how much it exceeded the similar amount in the preceding year), but it appears to have been on the order of $30 thousand. Adding that amount to the 1977 surplus and subtracting it from the 1978 surplus gives a decline from $186 to $152 thousand. (See Table 1.)

Receipts from dues and subscriptions rose from $744 thousand in 1977 to $903 thousand in 1978. For the reason just given, the true rise was perhaps $60 thousand smaller. Even so, it was on the order of $100 thousand. About $35 thousand of the rise came from the increase in rates for dues and subscriptions of 5 percent, which went into effect at the beginning of 1978. The number of subscribers rose 5.5 percent, accounting for another $15 thousand gain. The remaining rise may be attributed to a 9.6 percent increase in the number of members, which was heavily concentrated in the category of members paying the highest dues. The increase in membership resulted from a campaign by the Secretary to induce former members to rejon so that they would be listed in the 1978 *Directory*.

Now that the *Directory* has been published, the proportion of members failing to renew can be expected to increase. But since most of those who rejoined to get into the *Directory* did so after the beginning of 1978, some of their payments will be reflected in the 1979 income statement, which is on an accrual basis. On balance we expect a modest reduction in receipts from dues and subscriptions in 1979 as shown on the income statement (and a slight gain after correction for inclusion in 1978 of dues applicable to 1977).

The proposed budget for 1979, approved by the Executive Committee at its meeting on March 16, shows modest increases in receipts and expenditures with a surplus of about the same size as that for 1978 after the adjustment previously described. The Managing Editors of the two journals and the Secretary are to be commended for holding down expenditures. The rise from 1977 to 1978 in total expenditures was 6.3 percent. The budgeted rise from 1978 to 1979 is only 4.4 percent.

Whereas 1969–75 was a period of persistent deficits which at one point reduced the net worth of the Association below zero, 1979 will be the fourth year with a large surplus. The turnaround resulted from raising and restructuring the dues schedule. In my last report, I recommended planning on the assumption that the surplus would diminish $50 thousand each year, so that a dues increase would be needed to go into effect on January 1, 1981. Since the surplus has not diminished at the expected rate, a dues increase will not be needed so soon unless the Association decides to expand its activities. If, contrary to expectation, the surpluses continue in the neighborhood of $150 thousand, consideration should be given to a dues reduction.

RENDIGS FELS, *Treasurer*

TABLE 1—AMERICAN ECONOMIC ASSOCIATION BUDGET, ACCOUNTING BASIS, 1979
(Thousands of dollars)

	1977 Actual (Audited)	1978 Budget (3-17-78)	1978 Actual (Audited)	1979 Budget (3-16-79)
REVENUE				
Operating Income				
Dues and subscriptions	744	795	903	880
Advertising	77	81	87	90
JOE subscriptions	19	20	16	20
Sales—Miscellaneous	36	36	33	65
Sales—Mailing list	34	35	41	40
Sales—*Index*	100	87	49	60
Annual meeting	–	–	1	–
Other income	32	20	20	20
Total Operating Income	1,042	1,074	1,150	1,175
Investment Income				
Interest and dividends	41	45	58	57
Real capital gains (losses)	19	8	(20)	(22)
Total Investment Income	60	53	38	35
TOTAL REVENUE	1,102	1,127	1,189	1,210
EXPENSES				
Publications				
American Economic Review	284	305	303	307
Journal of Economic Literature	297	322	338	353
Directory	50	50	50	55
Job Openings for Economists	27	27	28	30
Index of Economic Articles	31	23	20	20
Total Publications Expenses	689	727	739	765
Operating and Administrative				
Salaries	103	110	107	114
Rent	9	10	10	10
Mailing list maintenance, etc.	28	36	38	38
Auditing and legal	17	9	9	11
Office supplies	11	12	11	12
Postage	14	16	15	18
Annual meeting	6	6	4	6
Federal income tax	13	15	12	13
Miscellaneous	18	17	30	27
Total Operating Expense	219	231	236	249
Committees				
Women	9	9	9	9
Minorities	9	10	11	9
Political Discrimination	1	10	1	9
Others	19	12	10	12
Total Committee Expenses	38	41	31	39
TOTAL EXPENSES	946	999	1,006	1,053
SURPLUS (Deficit)	156	128	182	157

Report of the Finance Committee*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 29, 1978, with costs and market values as of that date. The total market value of the securities portfolio at year-end was $897,335. After making adjustments for cash additions and withdrawals, (including the addition of $150,000 in Association funds for permanent investment), we estimate that the Association's investment portfolio experienced a total investment return of +4.4 percent during 1978. This, unfortunately, is of course less than the rate of general price inflation in 1978.

The $897,335 total includes the funds remaining from a Special Grant made by the Ford Foundation in January of 1969 and subsequently commingled with the Association's account. As of December 29, 1978, the Association's portion of the aggregate account was $844,594 or 94.1 percent, and the Special Grant represented the remaining $52,741 or 5.9 percent of the total.

As reported last year, the Finance Committee had previously directed that the Association's investment portfolio hold a combination of both common stocks and fixed-income investments; and that the commitment to common stocks move in a range of 50 to 67 percent of the Association's combined funds. Further, the Finance Committee removed its previous maturity restriction on the investment of nonequity funds, authorizing the investment advisor to extend maturities when appropriate to take advantage of changes in the yield curve, especially in the intermediate-term sector.

In view of this policy and also as a result of the $150,000 addition to the portfolio, several investment changes were made during the year. Sales were made in Weyerhaeuser, J. C. Penney and Federated Department Stores; new purchases were made in Continental Illinois Corporation, Corning Glass, John Deere, Ocean Drilling, Fort Howard Paper, Proctor & Gamble, and Standard Oil of Ohio. Furthermore, additions were made in a number of issues already held. Finally, a modest extension of bond maturities was made by moving some funds into the one- to two-year sector in order to take advantage of the recent inversion in the yield curve while avoiding the principal fluctuation risk associated with holding very long-term securities in a period of rising interest rates.

In terms of the portfolio's investment experience, the Committee can report that the +4.4 percent total return was about in line with the average returns of the widely followed market averages. (The Dow Jones Industrial Average had a total return of +2.6 percent.) When this result is combined with last year, a period of generally falling securities prices, the Association has a two-year total return of +2.6 percent as compared with the Dow Jones Average which experienced a return of −10.9 percent.

The Finance Committee anticipates that 1979 will be a period of slowdown in economic growth perhaps by some views resulting in a modest recession. This reduction in economic activity may be accompanied by some moderation in inflation and interest rates later in the year which might be greeted favorably by the securities markets. It was the Committee's view that the present low level of equity prices reflects a full discount for a number of uncertainties, and that common stocks are at nearly their lowest prices relative to earnings, cash flow, and asset value in the postwar period. Therefore, the Committee concluded that a continued meaningful exposure in the Association's portfolio to common stocks remains appropriate.

At its December 1978 meeting, the Finance Committee decided therefore to increase the flexibility to add to the Association's equity exposure by raising the combined portfolio's maximum equity limit to 75 percent from the previous 67 percent, thus setting an equity

*The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which follows.

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 29, 1978

	Value	Percent	Estimated Income	Estimated Current Yield
Cash Equivalents and Short-Term Securities	$187,218	20.9	$16,095	8.6
Medium-Term Securities	0	0.0	0	0.0
Long-Term Securities and Preferred Stocks	0	0.0	0	0.0
Convertible Securities	0	0.0	0	0.0
Equity Securities	710,117	79.1	28,172	4.0
TOTAL	$897,335	100.0	$44,267	4.9

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 29, 1978

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
Cash Equivalents and Short-Term Securities (20.9 percent)						
CASH EQUIVALENTS (0–1 YEAR) (2 percent)						
Cash			$18		$18	$2
Stein Roe Cash Reserves, Inc.	1981	1	1977	1	1976*	180
Subtotal Cash Equivalents			1995		1994	182
OTHER SHORT-TERM SECURITIES (1–5 YEARS) (20.6 percent)						
Northrn TR CAP NT (6.75 03/01/80)	40,000	96	38,281	97	38,850	2,700
US Treasury Notes (8.50 07/31/80)	40,000	97	38,988	100	39,921	3,400
US Treasury Notes (8 625 09/30/80)	40,000	98	39,038	100	39,982	3,450
US Treasury Notes (8.875 10/31/80)	30,000	98	29,428	100	29,965	2,663
US Treasury Notes (9.25 11/30/80)	40,000	99	39,488	100	39,934	3,700
	190,000		185,223		188,652	15,913
Subtotal Other Short-Term Securities			185,223		188,652	15,913
Total Cash and Fixed Income Securities			187,218		190,646	16,095
Equity Securities (79.1 percent)						
Utilities (4 3 percent)						
Central and Southwest	2,000	15	30,750	12	24,556*	2,680
Banks (7 8 percent)						
Continental Illinois	1,000	26	26,125	30	30,490	1,440
First Bank System	800	37	29,600	25	20,320*	1,536
			55,725		50,810	2,976
Other Financial (3 9 percent)						
Alexander and Alexander	1,000	28	27,500	9	9,325*	1,100
Foods and Related (7 7 percent)						
Philip Morris	400	71	28,200	44	17,726	820
Proctor & Gamble	300	89	26,663	85	25,523	900
			54,863		43,249	1,720
Paper and Forest Products (3 8 percent)						
Fort Howard Paper	700	39	26,950	41	28,791*	756
Machinery and Construction (4.9 percent)						
Deere	1,000	35	34,626	33	32,554*	1,500
Energy (15.4 percent)						
Cities Service	600	54	32,326	51	30,773*	1,920
Continental Oil	800	28	22,500	19	15,580*	1,200
Gulf Oil	800	23	18,700	17	13,321	1,520
Mapco	500	29	14,438	18	8,855	650
Standard Oil Ohio	500	43	21,250	38	18,848	440
			109,214		87,377	5,730
Oil Service (7.7 percent)						
Halliburton	400	66	26,400	63	25,310	720
Ocean Drilling & Explor	806	36	28,613	43	34,479	403
			55,013		59,789	1,123

Table 2 (*Continued*)

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
Drugs and Medical (9.5 percent)						
Abbott Lab.	1,000	34	33,750	21	21,360[a]	840
Merck	500	68	33,813	57	28,402[a]	950
			67,563		49,762	1,790
Electrical Products (4.6 percent)						
General Electric	690	47	32,517	36	24,536[a]	1,794
Computers and Office Equipment (5.0 percent)						
IBM	120	299	35,820	111	13,325[a]	1,651
Broadcasting and Publishing (3.6 percent)						
CBS	500	51	25,375	37	18,662[a]	1,300
Miscellaneous (21.7 percent)						
Corning Glass	400	53	21,300	60	23,984	752
Disney	800	40	32,100	25	20,161[a]	384
Eastman Kodak	600	59	35,176	65	38,911[a]	1,500
McDonalds	600	46	27,750	48	28,669[a]	216
Minnesota Mining and Mfg	600	63	37,875	54	32,473[a]	1,200
			154,201		144,198	4,052
TOTAL EQUITY SECURITIES (100 percent)			710,117		586,934	28,172
TOTAL SECURITIES AND CASH			897,335		777,580	44,267

[a]More than one cost basis.

range of 50–75 percent. In addition in order to facilitate the implementation of this investment policy for the Association's combined assets, it was voted to combine the existing temporary investment funds with the permanent portfolio while instructing the Treasurer to specify an amount to be kept in liquid assets to cover the needs of the Association.

ROBERT EISNER, *Chair*

Auditors' Report

To the Executive Committee of
The American Economic Association:

We have examined the statement of assets and liabilities of THE AMERICAN ECONOMIC ASSOCIATION (a District of Columbia corporation, not for profit) as of December 31, 1978 and 1977, and the related statements of revenues and expenses, changes in general and restricted fund balances and changes in assets and liabilities for the years then ended. Our examination was made in accordance with generally accepted auditing standards, and accordingly included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the accompanying financial statements present fairly the assets and liabilities of The American Economic Association as of December 31, 1978 and 1977, and its revenues and expenses, changes in fund balances and the changes in its assets and liabilities for the years then ended, in conformity with generally accepted accounting principles consistently applied during the periods.

Arthur Andersen & Co.

Nashville, Tennessee
February 19, 1979

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF ASSETS AND LIABILITIES
DECEMBER 31, 1978 AND 1977

Assets	1978	1977
CASH	**$ 87,860**	**$ 121,831**
INVESTMENTS, at market (Notes 1 and 2):		
Temporary investments	431,092	269,673
Permanent investments	902,093	735,327
	1,333,185	**1,005,000**
ACCOUNTS RECEIVABLE:		
Advertising, back issues, etc.	111,886	103,085
Allowance for doubtful accounts	(2,318)	(1,190)
	109,568	**101,895**
INVENTORY OF *Index of Economic Articles*, at cost	**52,837**	**31,072**
PREPAID EXPENSES	**17,056**	**3,583**
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $8,508 in 1978 and $6,992 in 1977	14,674	7,737
Total Assets	**$1,615,180**	**$1,271,118**

Liabilities and Fund Balances	1978	1977
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	**$ 313,060**	**$ 228,285**
DEFERRED INCOME (Note 1):		
Life membership dues	62,790	65,412
Other membership dues	330,986	191,337
Subscriptions	175,973	144,668
JOE	12,756	9,525
	582,505	**410,942**
ACCRUAL FOR DIRECTORY (Note 1)	**30,145**	**150,000**
FUND BALANCES:		
Restricted (Note 4)	58,774	53,111
Add (deduct)—Unrecognized change in market value of investments (Notes 1 and 3)	(2,584)	(2,419)
	56,190	50,692
General	674,494	451,602
Add (deduct)—Unrecognized change in market value of investments (Notes 1 and 3)	(41,214)	(20,403)
General fund-net worth	633,280	431,199
Total fund balances	733,268	504,713
Add (deduct)—Unrecognized change in market value of investments (Notes 1 and 3)	(43,798)	(22,822)
Net fund balance	**689,470**	**481,891**
Total Liabilities and Fund Balances	**$1,615,180**	**$1,271,118**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF REVENUES AND EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1978 AND 1977

	1978	1977
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$ 612,504	$ 481,370
Nonmember subscriptions	289,947	262,224
Job Openings for Economists subscriptions	16,021	19,325
Advertising	87,350	76,695
Sale of *Index of Economic Articles*	48,980	100,254
Sale of copies, republications, and handbooks	33,411	36,212
Sale of mailing list	41,240	33,850
Annual meeting	727	–
Sundry	20,099	32,049
	1,150,279	1,041,979
INVESTMENT GAINS (Note 2)	38,303	60,373
Net revenues	**1,188,582**	**1,102,352**
PUBLICATION EXPENSES:		
American Economic Review	302,640	221,025
Journal of Economic Literature	338,284	297,280
Papers and Proceedings	–	62,730
Directory publiction (Note 1)	50,000	50,334
Job Openings for Economists	28,413	26,920
Index of Economic Articles	19,983	31,405
	739,320	689,694
OPERATING AND ADMINISTRATIVE EXPENSES		
General and administrative—		
Salaries	106,738	102,664
Rent	10,086	8,721
Other (Exhibit I)	103,147	88,420
Committees	30,830	37,975
Annual meeting	3,830	5,816
Provision for federal income taxes (Note 6)	12,400	13,000
	267,031	256,596
Total expenses	**1,006,351**	**946,290**
REVENUES IN EXCESS OF EXPENSES	**$ 182,231**	**$ 156,062**

The accompanying notes to financial statements and Exhibit I are an integral part of this statement

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN GENERAL FUND BALANCE FOR THE YEARS ENDED DECEMBER 31, 1978 AND 1977

	Total	Operations	Market Value Adjustments
Balance at January 1, 1977	**$277,198**	**$ 28,095**	**$249,103**
Add—market value adjustments resulting from inflation (Note 1)	18,342	–	18,342
Add—revenues in excess of expenses	156,062	156,062	–
Balance at December 31, 1977	**451,602**	**184,157**	**267,445**
Add—market value adjustments resulting from inflation (Note 1)	40,661	–	40,661
Add—revenues in excess of expenses	182,231	182,231	–
Balance at December 31, 1978	**$674,494**	**$366,388**	**$308,106**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1978

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains (Note 4)	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics	$47,305	$ –	$ –	$3,416	$50,721
The Alfred P. Sloan Foundation, Chase Manhattan Bank, and Ford Foundation grants for increase of educational opportunities for minority students in economics	599	57,810	(58,409)	–	–
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	–	5,915	(5,915)	–	–
The Asia Foundation grant for Asian economists' membership dues to the American Economic Association and related travel expenses	616	–	(150)	–	466
The Carnegie Foundation grant for the Committee on the Status of Women in the Economics Profession	–	–	–	–	–
The National Science Foundation grant for support of a joint *U.S.-USSR* Symposium on the Economics of Technological Progress	–	–	–	–	–
The Minority scholarship fund for minority students applying for graduate work in economics	5,000	–	–	–	5,000
The Ford Foundation grant for development of a consortium on graduate studies in economics for minorities	(1,220)	10,000	(7,104)	–	1,676
The International Communication Agency grant arranging for a delegation of American economists to participate in a joint symposium with the Soviet Federation of Economic Institutions		14,219	(14,219)	–	–
Sundry	811	100	–	–	911
	$53,111	**$88,044**	**$(85,797)**	**$3,416**	**$58,774**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1977

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains (Note 4)	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics	$40,744	$ 2,101	$ –	$4,460	$47,305
The Alfred P. Sloan Foundation, Chase Manhattan Bank and Ford Foundation grants for increase of educational opportunities for minority students in economics	–	57,123	(56,524)	–	599
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	1,068	1,121	(2,189)	–	–
The Asia Foundation grant for Asian economists' membership dues to The American Economic Association and related travel expenses	1,067	–	(451)	–	616
The Carnegie Foundation grant for the Committee on the Status of Women in the Economics Profession	5	–	(5)	–	–
The National Science Foundation grant for support of a joint *U.S.–USSR* Symposium on the Economics of Technological Progress	-	706	(706)	–	–
The Minority scholarship fund for minority students applying for graduate work in economics	5,000	–	–	–	5,000
The Ford Foundation grant for development of a consortium on graduate studies in economics for minorities	–	–	(1,220)	–	(1,220)
Sundry	711	100	–	–	811
	$48,595	**$61,151**	**$(61,095)**	**$4,460**	**$53,111**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN ASSETS AND LIABILITIES FOR THE YEARS ENDED DECEMBER 31, 1978 AND 1977

	1978	1977
Cash, beginning of year	**$121,831**	**$ 46,373**
SOURCE (USE) OF FUNDS:		
Revenues in excess of expenses	182,231	156,062
Add- noncash charges-		
Depreciation	1,231	1,117
Directory publication (Note 1)	50,000	50,000
Market value adjustments (Note 1)	22,397	(19,267)
Funds provided by operations	255,859	187,912
(Increase) decrease in-		
Receivables and prepaid expense	(21,146)	68,999
Inventory of *Index of Economic Articles*	(21,765)	15,026
Investments	(328,185)	(108,153)
Office furniture and equipment	(8,168)	(1,827)
Increase (decrease) in-		
Accounts payable and accrued liabilities	(85,080)	60,270
Deferred income	171,563	(120,960)
Restricted funds	5,663	4,516
General fund, market value adjustment	40,661	18,342
Unrecognized change in market value of investments	(43,373)	(48,667)
Cash, end of year	**$ 87,860**	**$121,831**

The accompanying notes to financial statements are an integral part of this statement.

Notes to Financial Statements

(1) Significant Accounting Policies

Investments:

The Association accounts for its investments on a market value basis. Under the method used by the Association to value investments, the change in market value of corporate stocks during the year, after adjusting for an inflation factor (8.3 percent in 1978 and 5.9 percent in 1977), is recognized in income over a three-year period. The change in market value of treasury bills, commercial paper, etc., is reflected currently in income. The changes in market value of investments are allocated to the general and restricted fund balances as appropriate.

Accrual for Directory.

Approximately every three to five years, the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was published during 1978 and distributed at no cost to the membership. In order to match more properly the publishing cost of this directory with revenue from membership dues, the Association provides $50,000 annually for estimated publishing costs which will reduce actual directory expense in the year of publication.

Deferred Income

Revenue from membership dues and subscriptions to the various periodicals of the Association are deferred when received; these amounts are then recognized as income as publications are mailed to the members and subscribers.

Revenue from life membership dues is recognized over the estimated average life of these members.

(2) Investments and Investment Income

The following is a summary of investments held by the Association at December 31:

	1978		1977	
	Cost	Market	Cost	Market
Treasury bills, commercial paper, etc.	$ 623,068	$ 623,068	$484,389	$ 484,389
Corporate stocks	586,968	710,117	401,947	520,611
	$1,210,036	$1,333,185	$886,336	$1,005,000

Investment gains (losses) recognized in income for the years ended December 31, were as follows:

	1978	1977
Treasury bills, commercial paper, etc		
Interest	$38,968	$31,088
Change in market value	–	–
	38,968	31,088
Corporate stocks—		
Cash dividends	21,732	10,018
Increase (decline) in market value recognized (Note 3)	(21,530)	21,552
	202	31,570
Less Investment gains (losses) allocated to restricted fund (Note 4)	867	2,285
Investment gains (losses) included in income	$38,303	$60,373

(3) Unrecognized Change in Market Value of Investments

As described more fully in Note 1, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affect 1978 and 1977 revenues, and the amount of these market value increases (declines) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change December 31	
	1978	1977	1979	1980	1978	1977
1975	$ –	$28,913	$ –	$ –	$ –	$ –
1976	8,100	8,100	–	–	–	8,099
1977	(15,461)	(15,461)	(15,461)	–	(15,461)	(30,921)
1978	(14,169)	–	(14,169)	(14,168)	(28,337)	–
	$(21,530)	$ 21,552	$(29,630)	$(14,168)	$(43,798)	$(22,822)

Included in the above unrecognized changes as of December 31, are increases (declines) of ($2,584) and ($2,419) in 1978 and 1977, respectively, which have been allocated to a restricted fund. The amounts allocated are based on the percentage of the Association's total stock portfolio owned by this restricted fund.

(4) Restricted Fund:

The Association sponsors the Economics Institute, an organization that provides orientation programs for foreign graduate students of economics. The Policy and Advisory Board which determines overall policies applicable to Economics Institute is appointed by the President of the Association. Economics Institute participates in the investment program of the Association and its share of investments is accounted for as restricted funds in the accompanying statement of assets and liabilities. Investment income and market value adjustments applicable to Economics Institute which were allocated to the restricted fund were as follows:

	1978	1977
Net investment gains (losses) (Note 2)	$ 867	$2,285
Market value adjustments arising from inflation	2,549	2,175
	$3,416	$4,460

(5) Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan Payments by the Association and participating employees are based on the employee's compensation Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was $14,762 and $14,550 for 1978 and 1977, respectively.

(6) The Association

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under section 501(c)(3) of the U.S Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose This "unrelated business income" includes income from advertising and the sale of mailing lists.

The Association has been determined to be an organization which is not a private foundation.

EXHIBIT 1—THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1978 AND 1977

	1978	1977
Mailing list file maintenance and periodic mailing expenses	$ 37,563	$28,075
Accounting and legal	9,200	17,450
Office supplies	11,102	11,179
Postage	15,497	14,136
Dues and subscription	3,405	2,906
Telephone	3,260	3,199
Investment counsel and custodian fees	3,513	2,624
President and president-elect expenses	4,490	1,765
Travel and entertainment	858	1,955
Depreciation (straight-line method)	1,231	1,117
Uncollectible receivables	5,822	–
Currency exchange charges	1,504	693
Insurance and miscellaneous	5,702	3,321
	$103,147	$88,420

Report of the Managing Editor
American Economic Review

The record of papers received in 1978 is shown in Table 1. The number of papers submitted this year is 649, compared with 690 in 1977. Submissions to the *Review* were at their highest (879) in 1970, and on average have been declining ever since. We printed 108 papers in 1978, 6 fewer than the preceding year. There is a backlog of 85 accepted papers, 28 will appear March 1979, and the remainder in June or September.

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1959-78

Year	Submitted	Published	Ratio of Published to Submitted
1959	279	48	17
1960	276	46	17
1961	305	47	.15
1962	273	46	.17
1963	329	46	.14
1964	431	67	16
1965	420	59	14
1966	451	62	14
1967	534	94	18
1968	637	93	15
1969	758	121	.16
1970	879	120	.14
1971	813	115	14
1972	714	143	.20
1973	758	111	15
1974	723	125	.17
1975	742	112	.15
1976	695	117	17
1977	690	114	.17
1978	649	108	.17

The backlog of unprocessed manuscripts has remained at six months. That is, as of December 31, nearly all manuscripts received before the preceding June 30 have been processed. The authors have received some type of decision. It is not possible to cut this delay time further.

The delay time may nevertheless account for the decline in the number of manuscripts submitted. I have discussed methods of speeding up the review process with members of the Board of Editors, with my assistants, and with the Executive Committee. Each proposal runs up against the problem of cost and satisfactory standard of performance. Because the *Review* is an association journal, papers cannot be rejected on the basis of a cursory examination. Moreover, letters of complaint must be treated seriously, and the journal must be accountable to potential authors for the decisions taken on their manuscripts. This implies delay.

Number and Subject Matter of Submitted and Printed Papers

As Table 2 shows, we printed 108 regular papers this year, 52 main articles, and 56 communications, notes, comments, and replies. The size of the *Review* is slightly smaller than last year, 1,035 pages, compared to 1,067 in 1977 and 1,035 in 1976.

Table 3 shows the distribution of manu-

TABLE 2—SUMMARY OF CONTENTS, 1977 AND 1978

	1977		1978	
	Number	Pages	Number	Pages
Articles	50	635	52	663
Shorter Papers, including Notes, Comments and Replies	64	334	56	299
Special Articles	1	13	1	2
Dissertations		23		24
Announcements and Notes		53		37
Index		9		10
TOTAL	115	1067	109	1035

TABLE 3—SUBJECT MATTER DISTRIBUTION OF SUBMITTED AND PUBLISHED MANUSCRIPTS IN 1978

	Submitted	Published
General Economics and General Equilibrium Theory	6	5
Micro-Economic Theory	87	18
Macro-Economic Theory	46	5
Welfare Theory and Social Choice	48	10
Economic History, History of Thought, Methodology	11	1
Economic Systems	19	5
Economic Growth, Development, Planning, Fluctuations	44	5
Economic Statistics and Quantitative Methods	31	1
Monetary and Financial Theory and Institutions	51	7
Fiscal Policy and Public Finance	38	3
International Economics	63	11
Administration, Business Finance	19	6
Industrial Organization	46	9
Agriculture, Natural Resources	11	–
Manpower, Labor, Population	90	12
Welfare Programs, Consumer Economics, Urban and Regional Economics	39	10
TOTAL	649	108

scripts, classified by subject matter. The most popular fields are microeconomics, labor, international economics, monetary, fiscal, and macro theory, and welfare economics. This has not changed.

Administration

The Board of Editors has taken on two responsibilities which have smoothed the functioning of my job. One is to serve in an appeals capacity, reading manuscripts whose authors challenge a referee's judgment. The second is to serve as referees on communications and comments that when accepted are published together with an author's reply. Except in unusual circumstances (such as computational or mathematical detail) a comment is never sent to the original author for refereeing.

The Board consists of eighteen members, chosen by the Managing Editor, with the approval of the Executive Committee of the Association. Their names are printed in every issue. The cooperation of members of the Board is very satisfactory.

In March 1978, six new members of the Board were appointed by the Executive Committee for three-year terms. They are: Rudiger Dornbusch, William Oakland, Richard Roll, A. Michael Spence, William S. Vickrey, and S. Y. Wu.

Six members of the Board will have completed their terms at the end of 1978. Irma Adelman, David P. Baron, Robert J Barro, Laurits R. Christensen, David Laidler, and Frank Stafford. I wish to thank them for their high professional standards, work, and cooperation.

Finally I should like to thank the continuing members of the Board: Albert Ando, Elizabeth Bailey, David Bradford, Martin Feldstein, F. M. Scherer, and Jerome Stein.

The Papers and Proceedings

A major change in my editorial responsibilities occurred this year with the transfer of the *Papers and Proceedings* to the office of the *Review*. The *Papers and Proceedings* issue had been the responsibility of the Secretary's office, and we took it on at their request. Our first issue appeared May 1978 The burden of this effort fell on the shoulders of Wilma St. John and James Hanson, and they carried it out with great resourcefulness and energy. Hopefully a number of bugs in our procedure will be eliminated with more experience. For example, one author with a named professorship was erroneously listed as coauthor with the long deceased donor of his chair. This might be an example of perish and publish.

Expenses—Printing and Mailing

As Table 4 indicates, total printing and mailing expenses came to $195,882 this year, with more than 25 percent going to the *Proceedings* issue. When we started the year, we had no firm idea of the costs of editing and

TABLE 4—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING: 1978 *AER*

Issues	Copies Printed	Pages		Cost		
		Net	Gross	Issue[b]	Reprints[c]	Total
March	27,463	245	288	$ 32,320.68	$ 382.05	$ 31,938.63
May	27,864	520	536	55,979.64	978.30	55,001.34
June	28,000	265	288	35,205 02	910 75	34,294 27
September	28,000	244	280	33,925.61	479.50	33,446.11
December[a]	27,500	298	336	40,671.65	687.65	39,984.00
Annual Misc.[d]						1,217.26
TOTAL	138,827	1,572	1,728	$198,102.60	$3,438.25	$195,881.61

[a]Estimate
[b]Includes allocated cost of preparing mailing list.
[c]Credit resulting from charges to authors for additional reprints
[d]Includes extra shipping charges and storage costs of back issues.

TABLE 5—ACTUAL AND BUDGETED EXPENDITURES, 1971–79

	Printing and Mailing	Office Expenses	Total
1971	$120,120	$43,524	$163,644
1972	107,196	44,473	151,669
1973	117,873	49,121	166,994
1974	139,502	58,396	197,898
1975	129,476	63,372	192,848
1976	139,300	67,130	206,430
1977	141,769	70,788	212,557
1978[a]	210,000[c]	94,539	304,539
1978[b]	195,882[c]	87,766	283,648
1979[a]	207,635[c]	97,563	305,198

[a]Budget
[b]Actual.
[c]Includes *Papers and Proceedings* issue

printing the *Proceedings*, and we overestimated our overall budget. (See Table 5.) The 1979 budget of $207,635 for printing and mailing and $97,563 for office expenses should be more accurate. The addition of the *Proceedings* issue to our overall responsibilities has raised our expenses considerably and has put an extra burden on our staff.

Acknowledgments

I should like to thank my associates for their cooperation and hard work: Wilma St. John for her fine work as assistant editor on the regular issues of the *Review* plus the *Papers and Proceedings*; my colleague James Hanson, who serves as coeditor of the *Papers and Proceedings*; Deborah Franklin, our editorial assistant, and Sandra Overton, our secretary.

The following graduate students worked for the *Review* as proofreaders and hunters of false proofs: John Boschen, George Briden, John Chilton, Marvin Goodfriend, Robert King, Phillip Kott, and Joel Scheraga.

The following served as editorial consultants in the screening of manuscripts:

J. Albrecht
F. E. Bloch
L. Blume
G. Borjas
K. Boyer
R. P. Braeutigam
Y. Braunstein
R. Deb
A. T. Denzau
G. Dorman
D. Epple
R. Falvey
G. Faulhaber
A. Feldman
R. D. Feldman
M. T. Flaherty
R. Forsythe
H. L. Gabel
S. Garber
J. Geweke
R. Gilbert
G. Goldstein
F. Gollop
J. Gordon
R. H. Gordon
W. Greene
D. Hanson
M. Harris
M. Hashimoto
G. Hildebrandt
R. J. Hodrick
C. Lieberman
R. Mackay
M. M. Murphy
J. D. Richardson
J. Roberts
M. Rosenzweig
T. Russell
J. Rutledge
A. Schotter
S. Shavell
C. Stone

A. Strickland
H. R. Varian
B. Wascow
R. Wilder
K. Wolpin
A. Zelenitz

In addition to the members of the Board and the editorial consultants, I have sought and received the assistance of a large number of economists during the year. I wish to thank them for their cooperation and high standards in reading and evaluating manuscripts. The following have assisted as referees:

G. Ackley
B. Aghevli
A. Alchian
P. Allen
J. Anderson
M. Arak
S. Arndt
R. H. Arnott
K. J. Arrow
O. Ashenfelter
A. J. Auerbach
C. Azariadis
C. Azzi
M. N. Baily
B. Balassa
R. E. Baldwin
G. Ballentine
P. K. Bardhan
R. Barlow
W. Barnett
J. Barron
Y. Barzel
R. N. Batra
W. Baumol
V. S. Bawa
M. Beckmann
L. Benham
G. Benston
T. Bergstrom
E. Berndt
T. J. Bertrand
R. Betancourt
G. O. Bierwag
J. Bilson
J. Bishop
S. Black
O. Blanchard
R. G. Bodkin
W. Bomberger
J. P. Bonin
C. F. J. Boonekamp
T. E. Borcherding
R. Boyer
R. Braeutigam
W. Brainard
W. H. Branson
R. A. Brecher
F. P. R. Brechling
H. Brems
D. W. Bromley
J. P. Brown
M. Brown
E. K. Browning
W. Buiter
E. Burmeister
S. Burness
P. Cagan
J. Callen
G. A. Calvo
C. D. Cambell
D. Capozza
G. Carliner
J. A. Carlson
K. Carlson
D. Carlton
F. Casas
J. Cassing
R. E. Caves
P. L Cheng
G. C. Chow
C. F. Christ
L. Christensen
C. Clark
P. B. Clark
C. Clotfelter
W. L. Coats
P. R. P. Coelho
D. V. Coes
J. Conlisk
P. J. Cook
R. Cooter
W. M. Corden
J. C. Cox
M. A. Crew
J. B. Crockett
J. Crotty
A. Cukierman
R. G. Cummings
R. Dansby
M. R. Darby
R. d'Arge
A. Deardoff
F. de Leeuw
E. Denison
M. Denny
P. A. Diamond
W. E. Diewert
A. K. Dixit
P. Doeringer
F. T. Dolbear, Jr.
M. P. Dooley
J. H. Duloy
E. Eaton
R. Ehrenberg
R. Eisner
J. W. Elliott
E. Elton
W. Enders
R. Engle
D. Epple
W. Ethier
R. Fair
E. Fama
G. Fane
H. S. Farber
G. Faulhaber
E. Feige
G. Feiger
A. Feldman
R. Findlay
P. C. Fishburn
A. Fisher
R. Flanagan
J. Flanders
F. Flatters
B. Fleisher
A. M. Freeman
R. Freeman
A. Friedlaender
J. Fried
B. Friedman
R. Froyen
E. Furubotn
M. Fuss
C. Futia
R. Gallman
G. C. Galster
A. Gandolfi
I. Garfinkel
H. Genberg
J. F. Giertz
I. Gillespie
L. Girton
M. Goldberg
A. Goldberger
S. M. Goldfeld
C. Goodhart
R. H. Gordon
R. J. Gordon
E. Gramlich
E. Greenberg
R. Grieson
J. Griffin
H. Grossman
S. Grossman
N. Hakansson
R. Hall
R. Halvorsen
K. Hamada
R. Hamada
M. Hamburger
D. Hamermesh
B. W. Hamilton
E. A. Hanushek
J. R. Haring
J. P. Harkness
M. Hartley
M. Harris
R. Hartman
J. Hause
R. Haveman
A. Havenner
G. A. Hawawini
J. Heckman
J. M. Heineke
J. Hekman
W. P. Heller
E. Helpmann
J. V. Henderson
O. Hochman
W. Holahan
R. Holbrook

C. C. Holt
D. Holthausen
M Honig
H. Hori
I. Horowitz
G. Horwich
C. Hulten
R. P. Inman
M. D. Intriligator
Y. M. Ioannides
N. J. Ireland
N. L. Jacob
D. Jaffee
T. Johnson
R. W. Jones
M. Jones-Lee
P. Jonson
D. Jorgenson
P. Joskow
M. Kamien
E. J. Kane
J H. Kareken
E Katz
M. Keeley
P B. Kenen
J Kesselman
M. S. Khaled
A. Khan
C. Khang
B. Klein
L Klein
T Koizumi
K J. Kopecky
R. Kormendi
M. Kosters
M E. Kreinin
A. Krueger
J. Kurien
R. J. Lampman
W Landes
L. B. Lave
E. Lazear
J. Ledyard
H. Leibenstein
S. LeRoy
H. Levin
H. Levy
T Lewis
C M. Lindsay
D. Logue
M. C. Lovell
R. E. B. Lucas
S. McCafferty
R. A. McCain
J. J. McCall
B. T. McCallum
R. McCulloch
L. McKenzie
R. McKinnon
L. Maccini
R. J. MacKay
J. MacKinnon
W. Magat
S. Maital
J. H. Makin
B. Malkiel
C. Mallar
R. B. Mancke
M. Manser
E. Mansfield
S. A. Marglin
J. Marshall
A. L. Marty
J. P. Mattila
W. Mayer
J. Mayshar
J. Medoff
J. R. Melvin
R. Meyer
R. T. Michael
P. Mieszkowski
M. A. Miles
M. Miller
D. E. Mills
J. J. Minarik
J. A. Mirrlees
T. Miyao
H. Mohring
J. Moore
J. B. Moore
T. G. Moore
S. Morley
J. Muellbauer
D. Mueller
A. H. Munnell
R. Muth
R. Myerson
E. Nadel
K. Nagatani
J. P. Neary
P. Neher
D. M. G. Newbery
J. Newhouse
P. K. Newman
J. Niehans
R. Noll
W. Nordhaus
J. B. Nugent
W. Oakland
W. E. Oates
R. Oaxaca
E. Olsen
J. Olson
M. Olson
J. A. Ordover
D. K. Osborne
M. Ott
T. Page
M. Paglin
J. Panzar
R. E. Park
M. Parkin
D. Parsons
M. V. Pauley
S. Pejovich
J. Peles
J. Pencavel
R. Perlman
S. Perrakis
M. Perry
J. Pettengill
N. Phelps
D. Pines
M. Piore
J. E. Pippenger
C. Plott
C. G. Plourde
M. Polinsky
W. Poole
J. Pomery
R. A. Posner
U. Possen
E. Prescott
I. Pressman
D. Purvis
B. Putnam
J. Quirk
R. Radner
R. H. Rasche
N. Rau
G. Rausser
A. Raviv
A. Razin
E. E. Reinhardt
S. Resnick
G. F. Rhodes
J. D. Richardson
J. Riedel
S. Robinson
J. D. Rodgers
C. A. Rodriguez
D. Roper
H. Rose
S. Rose-Ackerman
S. Rosefielde
S. Rosen
M. Rosenzweig
J. Rosse
S. Rottenberg
J. C. R. Rowley
R. Ruggles
W. R. Russel
E. Sadka
L. Sahling
J. Salop
S. Salop
A. Santomero
R. Saposnik
R. Sato
T. Saving
B. Schiller
R. Schmalensee
R. Schuler
M. Schupack
A. J. Schwartz
N. Schwartz
R. Schwartz
D. Schwartzman
W. Schwert
A. K. Sen
R. S. Seneca
E. Seskin
P. Shapiro
R. Sherman
R. J. Shiller
R. Shishko
B. Shitovitz
C. D. Siebert
E. Silberberg
H. A. Simon
D. Sjoquist
S. Slutsky
K. Small
J. P. Smith

R. Smith
V. L. Smith
E. Smolensky
W. Springer
T. N. Srinivasan
D. A. Starret
M. Stewart
H. Stokes
M. Strober
J. D. Stryker
D. Suits
A. A. Summers
J. E. Tanner
G. Tauchen
J. Taylor
R. L. Teigen
L. Telser
L. Thurow
N. Tideman
R. D. Tollison
E. Tower
R. Townsend
S. C. Tsiang
H. Tuckman
S. Turnovsky
D. Usher
A. Vanags
J. Vanek
W. S. Vickrey
D. R. Vining, Jr.
W. Vroman
P. Wachtel
M. L. Wachter
M. Ward
J. Warner
F. Warren-Boulton
C. Watts
H. W. Watts
W. E. Weber
B. Weisbrod
L. W. Weiss
R. Weiss
Y. Weiss
M. L. Weitzman
F. Welch
F. W. Westfield
J. Whalley
W. C. Wheaton
A. Whinston
L. J. White
G. A. Whitmore
D. Wildasin
G. Wilensky
D. S. Wilford
J. G. Williamson
J. P. Williamson
R. Willig
R. J. Willis
R. E. Wong
G. Wood
J. H. Wood
J. Yellen
G. Yohe
P. A. Yotopoulos
E. E. Zajac
R. Zeckhauser

GEORGE H. BORTS, *Managing Editor*

Report of the Managing Editor
Journal of Economic Literature

This report, originally prepared for the August 1978 meetings at Chicago, has been revised at the year's end.

Table 1 illustrates the projected allocation of space in the *Journal of Economic Literature* for 1978 as well as the comparisons for the years 1974 through 1977. Table 2 classifies the material by subject both for the 1978 issues and for the totals for the period 1969–78. And, finally, Table 3 classifies the material by technical difficulty.

Members will note that we published four survey articles during 1978 and that we published six articles on the literature in subfields. We also published one article analyzing bibliographic development, as such; since it deals in some major sense with the editorial policy of the *Journal*, I draw your attention to it.

We have, in various stages of completion, commissioned survey articles on economic education, the welfare implications of national income accounting, economics of information and uncertainty, and the economics of advertising. Essays on the literature relating to the theory of the firm, revisiting economic development, methodology (two), and the development of national accounts are also scheduled.

During 1978 we have published the annual *Index* for 1974. During 1979 we plan to produce the annual *Indexes* for 1975 and 1976. The manuscript for 1977 will be sent to press during 1979.

Circulation of the annual *Indexes* has been something of a puzzler. We have compared the price of these volumes with the prices for similar volumes in the field of political science and the actual price of our volume is significantly lower, as well as a lower price per entry. Nonetheless, demand for these volumes has been less than we had anticipated. Coverage in these volumes will be as formerly except that we have increased slightly the number of journals covered.

The Chancellor and the Dean of the Faculty of Arts and Sciences of the University of Pittsburgh have again allocated some University of Pittsburgh support to the *Journal*. Their willingness to do so, particularly in

TABLE 1—QUANTITATIVE ANALYSIS OF *JEL* CONTENTS, 1974–78
(Number of pages in brackets)

	1974		1975		1976		1977		1978	
	No.	Pages	No	Pages	No	Pages	No	Pages	No	Pages
Survey articles	3	(102)	3	(119)	3	(116)	4	(127)	4	(180)
Essays on subfields	5	(99)	5	(100)	8	(157)	4	(99)	4	(79)
Review articles	1	(5)	-	–	–	-	1	(9)	1	(12)
Articles about economic literature	–	–	1	(11)		–	–	–	2	(39)
Communications	13	(71)	12	(36)	2	(7)	15	(62)	1	(3)
Books annotated	1211	(229)	1203	(223)	1204	(253)	1212	(246)	1200	(259)
Books reviewed	168	(239)	183	(282)	185	(278)	172	(274)	182	(286)
Journal issues listed and indexed	986	(180)	908	(177)	901	(159)	970	(174)	921	(180)
Number of individual articles	7360	–	6788	–	6211	–	7164	–	7344	-
Subject index of journal articles	-	(338)	–	(349)	–	(328)	–	(329)	–	(360)
Abstracts of articles	1645	(312)	1637	(331)	1502	(309)	1589	(326)	1649	(338)
Total pages*		(1671)		(1700)		(1664)		(1713)		(1873)

*Includes, in addition to listed pages, classification systems, table of contents, indices, journal subscription information, etc.

TABLE 2—CLASSIFICATION BY SUBJECT, 1969–78

	1978 Commissioned Survey	1978 Creative Curmudgeon Essays	1969–78 All Articles Total[a]
01 General	–	2	8
02 Theory	1	2	26
03 Thought (Methodology)	–	2	23
04 Economic History	–	1	4
05 Comparative Systems	–	–	4
11–12 Growth and Development	–	–	6
13 Stabilization	–	–	2
21–22 Econometric, Statistical Theory, Statistics	–	–	3
31 Monetary Economics	–	–	7
32 Fiscal Economics	1	–	6
40–44 International Economics	–	–	11
50 Managerial Economics	–	–	1
60 Industrial Organization, Industrial Regulation	–	–	1
70 Agricultural and Resource Economics	–	–	2
80 Labor Economics	2	–	8
90 Applied Welfare Economics, Regional Economics	–	–	7
TOTALS	4	7	119

[a]Includes all review articles on books, general essays on all literature.

TABLE 3—CLASSIFICATION BY TECHNICAL DIFFICULTY, 1969–78

	1978 Surveys	1978 Creative Curmudgeon Articles	1969–78 Totals: Surveys, Creative Curmudgeon Articles; Others[a]
Most Difficult	–	1	21
Some Difficulty	2	1	53
Not Difficult	2	5	45
TOTALS	4	7	119

[a]Review articles or books and general essays on all literature; excludes very short communications.

this period of severe retrenchment and particularly tight budgets, illustrates an understanding of and a devotion to scholarly work in the economics discipline. I thank them frequently privately and I take this opportunity again to do so publicly.

Four members of the Board of Editors have completed their terms. I wish to convey to them publicly (although I have already done so privately) my great appreciation of their tremendous help. George R. Feiwel, University of Tennessee; Michio Morishima, London School of Economics; James K Kindahl, University of Massachusetts, Charles Z. Wilson, Jr., University of California-Los Angeles have been most helpful colleagues. The other members of my Board who will continue for one or two years have been similarly cooperative; I look forward to our continuing association. I have nominated four individuals to replace the departing people.

I also wish to thank the following economists (plus three who have chosen to remain anonymous) for advice and assistance in the commissioning, refereeing, and revising of articles:

Moses Abramovitz
Abram Bergson
Mark Blaug
Paul Davidson
Harry Oshima
John Sheahan
George J. Stigler
Martin L. Weitzman
Gordon C. Winston

Finally, each of the members of the *Journal* staff have done their usual superb work during this year: the associate editor, Naomi Perlman; the assistant editor, Drucilla Ekwurzel; the principal secretary, Lyndis Rankin; the clerk for the quarterly and annual indexing, Margaret Yanchosek; and my administrative assistant (who works half-time for this *Journal*), Robin Bandemer.

MARK PERLMAN, *Managing Editor*

Report of the Director
Job Openings for Economists

During 1978, employers advertised 1,647 new vacancies. Of these, 1,150 (70 percent) were classified as academic and 497 (30 percent) were nonacademic. Last year, employers advertised 1,470 new vacancies; 68 percent were academic and 32 percent were nonacademic. The division between academic and nonacademic remained roughly the same, but the total number of new vacancies increased by 12 percent. Table 1 shows the total listings (employers), total vacancies advertised, new listings and new vacancies by type for each issue of *JOE* in 1978.

TABLE 1—JOB LISTINGS FOR 1978

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	105	220	80	155
April	71	121	61	92
June	39	64	34	52
August	52	117	49	110
October	100	245	89	212
November	97	285	97	285
December	168	432	91	244
Subtotals	632	1,484	501	1,150
Nonacademic				
February	18	55	15	41
April	17	59	15	50
June	24	82	23	81
August	16	60	12	46
October	31	112	30	107
November	20	97	20	97
December	27	139	15	75
Subtotals	153	604	130	497
TOTALS	785	2,088	631	1,647

Universities with graduate programs and four-year colleges continue to be the major source of job listings. They constituted 44 and 37 percent, respectively, of total employers. This is comparable to last year's 46 and 32 percent for the two. Table 2 shows the number of employers by type for each 1978 issue. The distribution is similar to that in 1976 and 1977.

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in mathematics and statistics appear to be the type of economists that employers are seeking. The applied area of specialty seems to be of secondary importance. Table 3 shows the number of citations by field classification during 1978. General economic theory (000) led, followed by monetary and fiscal (300), econometrics and statistics (200), welfare and urban (900), and business administration (500).

The proposed 1979 budget and the 1978 (adopted and actual) and 1977 (adopted and actual) budgets are given in Table 4. The 1978 approved budget for *JOE* projected a deficit (including allocated costs) of $7 thousand. The estimated actual deficit is $11.6 thousand. Total revenues are expected to be $16,300, total direct costs $12 thousand, and total indirect costs $16,100. If indirect costs are excluded, *JOE* continues to be self-financ-

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1978

Issue	Four-Year Colleges	Universities with Graduate Programs	Junior Colleges	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	45	59	–	8	2	1	1	6	1	123
April	40	30	–	5	1	1	1	7	3	88
June	21	18	–	5	1	3	1	12	2	63
August	19	33	–	4	1	1	2	6	2	68
October	41	59	–	9	4	2	3	11	2	131
November	43	54	–	9	2	1	2	4	2	117
December	78	90	–	17	–	2	1	6	1	195
TOTALS	287	343	–	57	11	11	11	52	13	785

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1978

Field[a]	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	91	72	29	47	110	113	187	649
Growth and Development (100)	17	19	15	21	34	27	57	190
Econometrics and Statistics (200)	38	39	22	26	53	47	84	309
Monetary and Fiscal (300)	42	36	24	31	64	45	119	361
International Economics (400)	16	16	14	19	32	27	63	187
Business Administration, Finance, Marketing and Accounting (500)	57	42	15	11	40	28	55	248
Industrial Organization (600)	29	33	14	17	35	41	58	227
Agriculture and Natural Resources (700)	19	25	16	17	29	21	48	175
Labor (800)	25	16	7	12	24	28	48	160
Welfare and Urban (900)	51	26	18	17	41	49	78	280
Related Disciplines (A00)	5	–	–	3	6	7	11	32
Administrative Positions (B00)	9	2	6	4	11	8	7	47
TOTALS	399	326	180	225	479	441	815	2,865

[a]Fields of specialization codes are from the *Journal of Economic Literature*

TABLE 4—JOB OPENINGS FOR ECONOMISTS
Budget for 1979 (in thousands)

	1979 (Proposed)	1978 (Adopted)	1978 (Estimated)	1977 (Adopted)	1977 (Actual)
Revenue					
Subscriptions			16 0		19.3
Miscellaneous			3		.7
Total Revenue	$20	$20	$16 3	$21	$20.0
Expenses					
Direct					
Computer	2		1.5		.8
Typewriter Rental	2		1 7		2.4
Postage	4		3.9		3.4
Printing	5		4.4		3 6
Salaries	–		0		.2
Miscellaneous	1		.3		.5
Total Direct	14		11.8		10 9
Indirect					
Salaries	16		16.1		16.1
Total Expenses	$30	$27	$27 9	$25	$27.0
SURPLUS (DEFICIT)	(10)	(7)	(11.6)	(4)	(7)

ing. Our auditors have opined that *JOE* is an "unrelated business" and profits are taxable. The application of commonly accepted accounting principles should be sufficient to keep *JOE* from showing a profit in 1979 and subsequent years. The proposed budget for *JOE* for next year projects revenues of $20 thousand, total direct costs of $14 thousand, and total indirect costs of $16 thousand. This leads to a projected accounting deficit of $10 thousand.

Violet Sikes continues to do virtually all the work involved in the publication and distribution of *JOE*—typing, editing, layout, subscriptions, production, and mailing. I am pleased to express my great appreciation to her.

C. ELTON HINSHAW, *Director*

The Committee on the Status of Women in the Economics Profession

In establishing the Committee on the Status of Women in the Economics Profession (CSWEP) in 1971, the American Economic Association recognized that women were not sufficiently represented in the economics profession and gave official sanction to efforts to increase the role and participation of women in economics. To this end, CSWEP has undertaken a number of activities aimed at increasing the number of women active in the profession and has attempted to monitor their role and activities in economics. Thus this report will briefly discuss CSWEP's activities and the status of women in the academic labor market for economists.

I. CSWEP Activities

CSWEP's activities aimed at increasing the participation of women in the economics profession fall into two major groups: those aimed at enhancing the workings of the labor market for economists, particularly with respect to women; and those aimed at increasing the visibility of women and women's issues in the economics profession.

In terms of informational activities, the maintenance of the roster and publication of the *Newsletter* are the most important. CSWEP maintains a roster of all women economists who have registered with it. Each listing on the roster states the highest degree earned, current job, and fields of interest. Thus potential employers can obtain from the roster a list of all potential women candidates who fit a given job description (for example economists specializing in money and banking with more than three years experience). Consequently the roster ensures that no potential woman candidate will be excluded from consideration for lack of information. The roster listings are sold at a modest fee and are widely used by academic departments, government agencies, and industry.

While the roster primarily serves the needs of potential employers to ensure that they have full information about the pool of women candidates, the list of jobs in the *Newsletter* serves the needs of potential candidates. Although much of the job information in the *Newsletter* is also in the Association's publication of *Job Openings for Economists*, it is felt that the additional listings in the *Newsletter* are worthwhile. This is particularly true for people who are not actively looking for a job, but might learn of a suitable opening through the *Newsletter* listings. The *Newsletter* also provides information about issues of concern to women economists.

In addition to trying to improve the workings of the job market by enhancing the flow of information, CSWEP has attempted to increase the participation of women in the economics profession by sponsoring sessions at the annual meetings of the Association. While these sessions do not exclusively focus on women's issues, they attempt to focus on topics that might be of concern to women and in which women could be expected to be working. In addition, since the papers in this session are contributed rather than invited, the CSWEP session provides an outlet for less established economists, either male or female, at the annual meeting of the Association.

CSWEP has recently extended its activities to the meetings of the regional associations. During the academic year 1978–79, CSWEP sessions will be held in the meetings of the Southern Economic Association, the Midwestern Economic Association, and the Eastern Economic Association. In this connection, CSWEP hopes to establish regional representatives who will plan the CSWEP program at each of the regional meetings and encourage the participation of women economists at these sessions.

At each of the meetings of the economic associations, CSWEP also holds an open meeting and maintains a hospitality suite as a means of encouraging people to exchange their concerns about the role of women in the economics profession and discuss alternative ways to improve the role of women in the profession.

In this connection, during the past year

CSWEP has taken an active role on behalf of those members of the Association who felt that it was inappropriate for the Association to hold its meetings in states that have not ratified the Equal Rights Amendment (ERA). To this end, CSWEP made presentations before the Executive Committee of the Association in December 1977 and March 1978, urging it to move from Atlanta the annual meeting of the Association to be held in December 1979.[1] In both cases, the Executive Committee refused to vote to change the meeting site, citing as reasons: 1) the nonpolitical charter of the Association; 2) the fact that such an act would constitute a secondary boycott; and 3) existing contractual obligations. The issue was raised again in the Association's Open Business Meeting, which was held at Chicago in August 1978, and it was narrowly defeated by the members of the Association who were present at that time.

II. The Role of Women in the Economics Profession

While CSWEP actively attempts to promote the participation and visibility of women in the economics profession, the status of women within the profession must ultimately depend upon the kinds of jobs and responsibilities undertaken by women economists. As a primarily academic profession, this is best measured by the distribution of women econo-

[1]In addition, CSWEP raised the question of moving the 1978 meetings from Chicago, but it was generally felt that contractual obligations made such a move impossible

TABLE 1—DISTRIBUTION OF FULL-TIME FACULTY BY TYPE OF INSTITUTION, ACADEMIC YEAR, 1977–78

	Chairman's Group			Other Ph D Departments			Only M A Departments			Only B A Departments		
		Female			Female			Female			Female	
	Total	Number	Percent	Total	Number	Percent	Total	Number	Percent	Total	Number	Percent
Existing												
Professor	614	8	1 3	525	10	1 9	166	7	4 2	255	12	4 7
Associate	231	6	2 5	377	18	4 8	143	5	3 4	287	18	6.2
Assistant	329	37	11 2	259	25	9 7	149	18	12 1	373	32	8.6
Instructor	129	4	3 1	45	7	15 6	30	9	30 0	41	15	16 5
Other	70	10	14 3	66	6	9 1	20	2	10 0	29	5	17.2
New Hires												
Professor	12	0	0 0	7	0	0 0	1	0	0 0	3	0	0 0
Associate	7	0	0 0	9	0	0 0	5	0	0.0	3	0	7 7
Assistant	71	8	11 3	62	9	14 5	78	4	8.3	74	8	10 8
Instructor	12	1	8 3	15	3	20 0	11	2	18 2	39	5	12 8
Other	9	1	11.1	10	3	30 0	1	0	0 0	4	2	50 0
Promotion to Rank (1976 77 to 1977–78)												
Professor	34	0	0 0	29	0	0 0	13	1	7 7	24	1	4 2
Associate	26	1	3 8	40	0	0 0	20	0	0 0	41	5	1.2
Assistant	5	0	0 0	6	0	0 0	7	2	28 6	16	1	6 3
Instructor	–	–	-	-	–	-	–	-	-	–	-	–
Other	–		-	–	–	-	-	-	–	-	–	–
Tenured at Rank (1976-77 to 1977–78)												
Professor	7	0	0.0	7	0	0 0	3	0	0 0	4	0	0 0
Associate	17	1	5 8	25	0	0.0	14	1	7.1	29	1	13.8
Assistant	2	0	0 0	11	1	9 1	3	1	33 1	18	3	16.7
Instructor	–	–	-	-	–	–	–	–	-	-	–	–
Other	–	–	–	–	–	–	–	–	-	-	–	–
Not Rehired (1976-77 to 1977-78)												
Professor	14	1	7.7	25	0	0.0	10	3	30 0	11	0	0.0
Associate	16	0	0 0	10	0	0 0	3	1	25.0	13	1	7.7
Assistant	32	3	9 4	27	1	3 7	23	6	26 1	49	8	16.3
Instructor	1	1	100 0	11	0	0.0	6	0	0 0	22	3	13.7
Other	16	2	12.5	11	0	0 0	0	0	0 0	3	0	0 0

mists among various types of academic institutions and the flow of young women economists into these institutions.[2]

Table 1 provides a summary of the distribution of academic jobs at the beginning of the academic year 1977–78 and the changes that took place between this and the previous year.[3] This table presents information in terms of four types of departments: the Chairman's group; other Ph.D. departments; M.A. departments; and B.A. departments. The Chairman's group consists of the forty-three departments that focus on research and the training of Ph.D.s in economics. In terms of stature, it is generally agreed that academic appointments at a department within the Chairman's group carry the most prestige. Thus this discussion will tend to focus upon the role of women in the Chairman's group as a bellwether for the entire economics profession. The other Ph.D. granting departments primarily focus on undergraduate education, but also have a viable Ph.D. program. The M.A. departments similarly have a primary focus upon undergraduate education, but also have a Master's program. Finally, the B.A. departments are exclusively concerned with undergraduate teaching.

According to Table 1, the existing participation of women in the academic side of the economics profession is distressingly small. Within the Chairman's group, there are only eight women who are full professors, six who are associate professors, thirty-seven who are assistant professors and four who are instructors, respectively representing 1.3 percent of the full professors, 2.5 percent of the associate professors, 11.2 percent of the assistant professors, and 3.1 percent of the instructors. Although the percentage of women in each category is slightly higher for the other departments, the figures of the Chairman's group are representative.

[2]In this issue Barbara Reagan has an interesting paper arguing that women economists appear to be subject to the "revolving-door syndrome" under which they are hired at junior levels but not retained at senior levels

[3]These figures and those of the subsequent tables are based upon the Universal Academic Questionnaire distributed to all department chairmen and tabulated by the Association.

Since tenured positions carry the most prestige within the profession it is useful to focus on them. In this connection, it is interesting to note that as of 1977–78 the departments within the Chairman's group apparently feel that there are only eight women whose research and publication records are sufficiently strong to merit their appointments as full professor and six women whose records are sufficiently strong to merit their appointment as associate professors. Moreover, during 1976–77, within the Chairman's group no woman was hired as a full professor, promoted to full professor, or hired as an associate professor. Although one woman within this group was promoted to associate professor and tenured at rank, one woman professor also left. Thus there appears to have been no net change in the stock of tenured women faculty members within the Chairman's group during 1976–77.

The situation with respect to the other departments appears to be equally bleak in 1976–77. During this year, no woman professor was newly hired by *any* economics department and only an associate professor was newly hired (by a B.A. department). One woman was promoted to professor in each of the M.A. and B.A. departments, while five women were promoted to associate professor by the B.A. departments. However, within the M.A. and B.A. departments, three woman professors were not rehired, and two associate professors were not rehired. Thus although the total stock of tenured women faculty appeared to grow in 1976–77, this growth can be called marginal at best.

In terms of changes that are occuring within the academic labor market, it is useful to consider the previous activity of those who were newly hired and the present activity of those who were not rehired. Table 2 indicates that within the Chairman's group, relatively more of the newly hired women were faculty at other institutions or were graduate students than their male counterparts. However, while 25.3 percent of the newly hired males in other Ph.D. departments came from other faculty positions, only 9.5 percent of the women hired by these departments held positions as faculty at other institutions. These latter figures are

TABLE 2—PREVIOUS ACTIVITY OF NEW HIRES AND CURRENT ACTIVITY OF THOSE NOT REHIRED BY TYPE OF INSTITUTION AND SEX, ACADEMIC YEAR, 1977–78

	Previous Activity of New Hires				Current Activity of Not-Rehired			
	Male		Female		Male		Female	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Chairman's Group	96	100.0	11	100.0	69	100.0	4	100.0
Faculty at Other Institutions	23	24.0	3	27 3	29	42.0	0	0.0
Graduate Student or Postdoctoral	54	56.2	7	63.6	2	2 9	0	0.0
Government	2	2 1	1	9.1	3	4.3	1	25.0
Business, Banking, Research	5	5.2	0	0 0	9	13.1	2	50.0
Other	12	12.5	0	0 0	26	37.7	1	25.0
Other Ph.D. Departments	95	100.0	21	100 0	94	100.0	3	100.0
Faculty at Other Institutions	24	25.3	2	9 5	29	30.9	2	66.7
Graduate Student or Postdoctoral	54	56.8	17	81.0	5	5 3	0	0.0
Government	3	3 2	0	0.0	14	14.9	0	0.0
Business, Banking, Research	40	4 2	0	0 0	16	17 0	1	33.3
Other	10	10.5	2	9 5	30	31.9	0	0.0
M A Departments	57	100.0	4	100.0	27	100.0	9	100 0
Faculty at Other Institutions	14	24 6	1	25.0	12	44.4	3	33.3
Graduate Student or Postdoctoral	35	61.3	3	75 0	2	7.4	0	0.0
Government	3	5.3	0	0 0	2	7.4	0	0.0
Business, Banking, Research	0	0.0	0	0.0	0	0.0	0	0 0
Other	5	8 8	0	0.0	11	40.8	6	66 7
B A Departments	146	100.0	26	100 0	87	100.0	14	100.0
Faculty at Other Institutions	60	34.2	5	19.3	31	35.6	4	28.6
Graduate Student or Postdoctoral	69	47.3	15	57 7	6	6 9	0	0.0
Government	6	4.1	0	0 0	6	6 9	1	7 1
Business, Banking, Research	14	9.6	3	11.5	14	16 1	4	28 6
Other	7	4.8	3	11.5	30	34 5	5	35.7

similar for the other departments and indicate that, on balance, men appear to have considerably more academic mobility than women. Stated alternatively, Table 2 indicates that men who are not rehired by one department largely remain in academia and take jobs at other academic institutions while women do not. This pattern is particularly striking with respect to the activities of those who are not rehired. Among men, the academic retention rate appears to have been about 40 percent among all departments and was 42 percent among the Chairman's group. Among women, the overall academic retention rate is somewhat less than 20 percent, and none of the women who left the Chairman's group took an academic job.

Thus Tables 1 and 2 indicate that the participation of women in the prestige jobs at the prestige institutions appears marginal at best. Equally unfortunate is the fact that relatively little change appears to be taking place with respect to women in this relatively select group. Consequently, although it is

TABLE 3—DISTRIBUTION OF SALARY FOR WOMEN FACULTY BY TYPE OF DEPARTMENT AND TIME IN RANK, ACADEMIC YEAR, 1977–78

Highest Degree Offered and Relative Salary for Rank	All Women		Time in Rank			
	Number	Percent	Total	Above Median	At Median	Below Median
All Departments	235	100.0				
Salary above median	89	37.9	100.0	58.4	30.3	11.3
Salary at median	78	33.2	100.0	6.4	80.8	12 8
Salary below median	68	28.9	100 0	19.1	16.2	64.7
Ph.D., Chairman's	61	100.0				
Salary above median	25	41.0	100.0	64.0	28.0	8.0
Salary at median	15	24.6	100.0	0.0	100.0	0.0
Salary below median	21	34 4	100.0	9.5	14.3	76.2
Ph.D., Other Departments	63	100.0				
Salary above median	29	46.0	100.0	62.1	27.6	10.3
Salary at median	17	27.0	100.0	5.9	70.6	23.5
Salary below median	17	27.0	100 0	23.5	0.0	76.5
M.A. Departments	29	100.0				
Salary above median	9	31.0	100.0	77.8	11.1	11.1
Salary at median	9	31 0	100.0	0.0	77 8	22 8
Salary below median	11	38.0	100.0	9.1	9 1	81.8
B.A. Departments	82	100.0				
Salary above median	26	31.7	100 0	42 3	42 3	15.4
Salary at median	37	45.1	100.0	10.8	78.4	10.8
Salary below median	19	23.2	100 0	31.6	36.8	31.6

important to note that women comprise relatively high proportions of the jobs at the junior academic ranks, unless these women begin to receive promotion and tenure, a negative demonstration effect may begin to take place. While women Ph.D.s are undoubtedly pleased to receive a junior faculty appointment at a department within the Chairman's group (or any other department for that matter), if a disproportionate number of these women fail to receive promotion and tenure, it is unlikely that women will perceive an academic career as being particularly attractive.

Nevertheless, it is important to note that the salary distribution of women faculty members appears to be in line with that of their male colleagues. This can be seen from Table 3, which gives the distribution of salary of women faculty by type of department and time in rank. In general, salary appears to be related to time in rank, with those whose time in rank is above the median having a salary that is above the median. Although data are lacking to compare the distribution of male and female salaries, Table 3 does not indicate the existence of any gross discrepancies between the distribution of salaries and the distribution of time in rank.

Ultimately, however, if more women are to play an active role in the economics profession, more women must be trained as economists. In this connection, Table 4 is interesting and indicates a rather sizable attrition rate between the granting of the B.A. degree and the Ph.D. degree. Specifically, in 1976–77, while 23.7 percent of all B.A. degrees were received by women, only 8.6 percent of Ph.D. degrees were received by women Although it takes four–five years to turn a B.A. into a Ph.D., these figures are quite representative of the past four or five years[4] and clearly indicate that relatively fewer women who receive economics training at the B.A. level choose to to on to graduate school and obtain a Ph.D. than do their male counterparts. Whatever the reasons for this decline, it is clear that the flow of new female Ph.D.s must be substantially increased if the

[4]See CSWEP reports in *Proceedings* issues 1974, 1975, 1976, 1977.

TABLE 4—DEGREES GRANTED IN ECONOMICS BY TYPE OF DEPARTMENT AND SEX, ACADEMIC YEAR, 1976–77

	All Departments	Ph.D. Departments Total	Ph.D. Departments Chairman's	Ph.D. Departments Other	M.A. Departments	B.A. Departments
Number Departments	343	94	43	51	39	210
Number Ph.D.s	628	628	408	225	–	–
Number female	54	54	33	21	–	–
Percent female	8.6	8.6	8.2	9.3	–	–
Number M.A.s	1,434	1249	610	1539	183	2
Number female	250	225	111	114	25	0
Percent female	17.5	18.0	18.2	7.9	13.7	0.0
Number B.A.s	10,759	5234	3196	2038	861	4664
Number female	2,547	1099	678	421	165	1283
Percent female	23.7	21.0	21.2	20.7	19.2	27.5
Number Other	877	347	36	311	147	383
Number female	176	37	7	30	1	138
Percent female	20.0	10.7	19.4	9.6	0.7	36.0

proportion of women in academic jobs is to increase substantially.

Table 5 provides information on the jobs taken by new Ph.D.s in 1976–77 and indicates that although relatively fewer women Ph.D.s took academic jobs overall than their male counterparts, at least among students in the Chairman's group, relatively more women took academic jobs than their male counterparts. This apparent discrepancy is explained by the fact that while 63.1 percent of new women Ph.D.s granted by the Chairman's group took academic jobs, only 13.5 percent of women Ph.D.s granted by other institutions took academic jobs. Since the Chairman's group is the primary source of Ph.D.s who

TABLE 5—DISTRIBUTION OF ACTIVITIES OF NEW PH D. DEGREES BY SEX AND TYPE OF DEPARTMENT, 1976–77

	All Ph.D. Departments		Chairman's Group		Other Ph D. Departments	
	Number	Percent	Number	Percent	Number	Percent
All Ph.D.s						
Total	714	100.0	414	100.0	300	100.0
Education	315	44.1	234	56.5	300	12.0
Government	71	9.9	35	8.5	81	27.0
Business, Banking, Research	96	13.4	47	11.4	49	16.3
Other	232	32.6	98	23.6	134	44.7
Male Ph.D.s						
Total	624	100.0	376	100.0	248	100.0
Education	284	45.6	210	55.9	74	29.8
Government	55	8.8	30	8.0	25	10.1
Business, Banking, Research	64	10.2	44	11.7	20	8.1
Other	221	35.4	92	24.4	129	52.0
Female Ph.D.s						
Total	90	100.0	38	100.0	52	100.0
Education	31	34.4	24	63.1	7	13.5
Government	16	17.8	5	13.2	11	21.2
Business, Banking, Research	32	35.6	3	7.9	29	55.7
Other	11	12.2	6	15.8	5	9.6

Table 6—Activities of 1976–77 Ph.D.'s by Type of Department and Sex

	Number of New Ph.D.'s Employed in:				
	Total	Education	Government	Business, Banking, Research	Other
All Ph.D. Departments					
Total	714	315	71	96	232
Number female	90	31	16	32	11
Percent female	12.6	9.8	22.5	33.3	4.7
Chairman's Group					
Total	414	234	35	47	98
Number female	38	24	5	3	6
Percent female	9.2	10.3	14.3	6.4	6.1
Other Ph.D. Departments					
Total	300	81	36	49	134
Number female	52	7	11	29	5
Percent female	17.3	8.6	30.6	59 2	3 7

take academic positions, it is encouraging to note that the bulk of new women Ph.D.s from these institutions entered the academic labor market. However, it is important to note that an annual flow of twenty-four women entering the academic labor market a year is not sufficient to change the distribution of academic positions held by women. This is shown by Table 6, which indicates that only 10.3 percent of the new Ph.D.'s from the Chairman's group taking academic jobs were women.

Although relatively fewer women who receive B.A. degrees in economics go on to

Table 7—Distribution of Ph D Student Support, by Type of Support, Sex, and Department, 1977–78

	All Ph.D. Departments		Chairman's Group		Other Ph.D. Departments	
	Number	Percent	Number	Percent	Number	Percent
All Students						
Total	3484	100.0	2290	100.0	1194	100 0
Tuition only	205	5.9	153	6.7	52	4.4
Stipend only	549	15.8	322	14.1	227	19.0
Tuition + stipend	1743	50.0	1264	55.2	479	40.1
No support	926	26.6	520	22.7	406	34 0
No record	61	1.7	31	1.3	30	2 5
Male Students						
Total	2945	100.0	1912	100 0	1033	100 0
Tuition only	163	5.5	120	6.3	43	4 2
Stipend only	484	16.4	280	14.7	204	19.7
Tuition + stipend	1469	49.9	1072	56.1	397	38.4
No support	792	26.9	429	22.4	363	35 2
No record	37	1.3	11	0 5	26	2.5
Female Students						
Total	539	100.0	378	100.0	161	100 0
Tuition only	42	7.8	33	8.7	9	5.6
Stipend only	65	12.1	42	11.1	23	14.3
Tuition + stipend	274	50.8	192	50.8	82	50.9
No support	134	24.9	91	24.1	43	26 7
No record	24	5.4	20	5.3	4	2.5

obtain a Ph.D. than their male counterparts, it is important to note that this is probably not due to lack of graduate student support on the part of the academic departments. This is shown clearly in Table 7, which indicates that although a slightly lower percentage of women students receive full support (tuition plus stipend) overall, the percentage of women students receiving some form of financial aid is virtually identical to that of male students. Thus it is likely that it is perceptions concerning their future status in the economics profession that makes women turn from graduate study in economics rather than a lack of financial support per se.

In conclusion then, although the economics profession and its related institutions have made a conscious effort to recruit and encourage women economists in recent years, it appears that, on balance, progress is still painfully slow. Women economists in academic institutions still comprise an extremely small percentage of the total, and the bulk of these women hold junior level, nontenured positions. Thus the real test of the commitment of the economics profession to enhance the status of women in its activities will occur in the next few years, when the presently nontenured women faculty come up for tenure and promotion. If a proportionate share of these women move up through the academic ranks, this will be a definite sign that the profession is serious about making women equal partners. If, however, a disproportionate share of young women economists are not retained, this will almost certainly be interpreted as a sign that the economics profession will remain an essentially male bastion. In this case, it is likely that able young women will increasingly turn away from economics and enter professions which they perceive will give them more attractive career opportunities.

ANN F. FRIEDLAENDER, *Chair*

Report of The Economics Institute's Policy and Advisory Board

During the summer of 1978 the Economics Institute held its twenty-first consecutive transitional training program for foreign students commencing graduate study in economics, agricultural economics, and business administration in *U.S.* universities. About 260 students took part in the 1978 summer program, the largest group ever to attend the Institute. Continued growth of the program attests to the need of many foreign students for the training in English, economic theory, mathematics and statistics that the Institute provides.

The Economics Institute now operates a substantial year-round program on the Boulder campus. The academic year program concentrates on intensive English training for students planning to enter *U.S.* graduate programs in economics and related fields. Some 75 students now take part in the fall term program, and some 100 in the spring term program. The academic year program assists students whose English proficiency requires more training than can be provided during the summer program. Many academic year students continue during the summer session. The academic year program has the side benefit of enabling the Institute to maintain a minimum year-round staff.

Many academic year students come to the Institute without prior admission to *U.S.* graduate programs. The Institute has therefore expanded its capacity to assist students in gaining university admission and in providing admitting institutions detailed professional evaluations of students' qualifications and needs.

Increasing difficulty in securing financial support from domestic sources has forced the Institute to move toward self-financing. Fund raising efforts are now concentrated on raising modest sums to provide scholarships to qualified students who cannot finance their stay at the Institute.

Foreign students now constitute about 37 percent of enrollments in M.A. programs and 29 percent of enrollments in Ph.D. programs in *U.S.* economics departments, and 44 percent of graduate enrollments in agricultural economics departments. *U.S.* graduate programs are providing a valuable service in advanced training of foreign students in economics and related subjects, and foreign students are a large part of total enrollments in many institutions. The Economics Institute is an important program in facilitating the transition to *U.S.* graduate study.

The Policy and Advisory Board met in Boulder during the summer of 1978. Members of the Board are Edwin Mills, Chair, John Day, Carlos Diaz-Alejandro, Carl Eicher, Anne Krueger, Axel Leijonhufvud, and Raymond Vernon.

EDWIN S. MILLS, *Chair*

Report of the Committee on U.S.–Soviet Exchanges

The fourth U.S.–Soviet economic symposium was held as scheduled from June 27–30 in Togliatti, an automobile manufacturing center on the Volga. The subject of the symposium was "Problems of Industrial Management." There were ten American participants: Abram Bergson, Harvard University; David Granick, University of Wisconsin (submitted a paper but did not attend because of transportation difficulties); Elton Hinshaw, Vanderbilt University and American Economic Association; David Kendrick, University of Texas; James March, Stanford University; Jesse Markham, Harvard University; Albert Rees, Princeton University; Lloyd Reynolds, Yale University; Richard Rosett, University of Chicago; Oliver Williamson, University of Pennsylvania.

Because of the pending trials of dissidents in the Soviet Union, and the decision of some *U.S.* scientists to cancel proposed visits there, the question was raised in May whether we should cancel the June meeting. After consulting with President Koopmans, we decided not to cancel, but rather to advise members of the delegation that they should follow their consciences on this point. One member of the delegation did withdraw on this account, while the others did not.

There were about thirty Soviet participants in the Togliatti meeting, mainly from the Institute of Economics, the Institute of World Economics and International Relations, the Institute of the U.S.A. and Canada, and the Economics Faculty of Moscow State University. About half of these were younger Soviet economists, some of whom had spent time in the United States and were fluent in English. One advantage of the meetings in the *USSR* is that it is possible to make contact with younger scholars who do not have enough seniority to be included in Soviet delegations to the United States.

As at previous meetings, we used the prepared papers mainly to generate roundtable discussion. The discussions were quite informative, and were also pleasingly straightforward and nonideological.

There was also some serious professional activity in Moscow: visits to Moscow State University and to several of the research institutes, including the Economic Research Institute of Gosplan. This is a large, prestigious and influential organization, and we spent an interesting afternoon discussing planning methodology. The remainder of the two-week visit was largely touristic: two days in Volgograd, two days in Baku, and two days in Leningrad.

Comparing 1978 with the previous *U.S.* visit in 1976, it seemed to us that the Soviet organizers have upgraded the status of this exchange, perhaps partly because it is now written into the protocol between the American Council of Learned Societies and the Academy of Sciences of the *USSR*. Our hotel and other accommodations were distinctly superior to those in 1976. The admission to Gosplan is probably significant, since this is not an easy organization to visit.

The Committee continues to feel that these annual visits are worth the effort and expense involved. The tentative plan, which we discussed with Academician Khatchatuov during our visit, is to hold the fifth symposium in the United States in June 1979, on the subject: "Long-term Structural Change in the American and Soviet Economies, 1930–1980." We should add that there is still no continuing source of finance for this exchange program. In each of the past three years the Committee has had to scramble for funds, each time from a different source. We hope to be successful on the next round, but we cannot be sure at this point.

We would like to express our appreciation to the International Communications Agency, which provided funds for the 1978 visit, and to the International Research and Exchange Agency (IREX), which provided valuable logistical support. We shall continue to consult with IREX and others on whether the present exchange program can be useful in stimulating longer-term individual research visits and collaborative research undertakings by *U.S.* and Soviet scholars.

LLOYD REYNOLDS, *Chair*
ABRAM BERGSON
JOHN MEYER

Report of the Representative to the National Bureau of Economic Research

Martin Feldstein, the new president of the National Bureau, appointed William Branson Program Director of the National Bureau's program of research in International Economics. Branson joins Robert Hall (Economic Fluctuations), Benjamin Friedman (Financial Markets and Monetary Economics), David Bradford (Business Taxation and Finance), and Richard Freeman (Labor Studies) as a director of one of the five major research programs at the National Bureau. National Bureau research also continued in Health Economics, Demography, Law and Economics, Social Insurance, Growth of the American Economy, and Business Cycles. Major studies on Youth Unemployment and Capital Formation were launched during the year.

Board of Directors: The tenure as members of the National Bureau's Board of Directors of Richard Bird, Wallace Campbell, and Wilson Newman ended during 1978. George Conklin, Stephan Kaliski, and Stephen Stamas were named to the Board of Directors and Charles E. McLure, Jr. was named Vice President.

Conferences and Workshops: During the summer of 1978 the National Bureau held a summer institute consisting of workshops in Labor Economics and in Business Taxation and Finance. Zvi Griliches organized a conference on "Econometrics of Panel Data and Self-Selection" as part of the workshop in Labor Economics. It is expected that similar workshops will be held during the summer of 1979 in Economic Fluctuations and in Macroeconomic Aspects of International Finance.

A conference of the Universities–National Bureau Committee was held in June 1978 on Low Income Labor Markets. A Universities–National Bureau Conference on the Economics of Information and Uncertainty, currently being planned by a committee chaired by John McCall, is expected to be held in 1979. The National Bureau will continue to hold conferences with scholars from universities, but the Universities–National Bureau Committee will no longer plan them. University scholars who have suggestions for topics for future conferences may send them to any university representative on the National Bureau Board (see the list in the National Bureau Annual Report) or to me at the Johns Hopkins University, Baltimore, Maryland 21218.

The Conference on Research in Income and Wealth did not meet during 1978, but an Income and Wealth Conference on National Income Accounting is scheduled for May 1979. The Conference on Income and Wealth series will continue.

A Conference on Commodity Markets, Models and Policies in Latin America was held in Lima, Peru in May, and a further Conference on Trade Prospects for the Americas is currently being planned for 1979. A workshop on Economics and Control Theory was organized by David Kendrick at Texas University, on May 24–26.

George de Menil and Robert Gordon organized the first National Bureau international seminar in macroeconomics, held in Paris on September 11–12, 1978. Victor Fuchs organized a conference in Palo Alto on January 27–28, 1978 on The Economics of Physician and Patient Behavior.

Publications: The following National Bureau books and conference volumes were published (by Ballinger Publishing Company, except where indicated) during 1978: *Alternatives for Growth: The Engineering and Economics of Natural Resource Development*, Harvey McMains and Lyle C. Wilcox, eds.; *Distribution of Economic Wellbeing*, F. Thomas Juster, ed.; "The Economics of Physician and Patient Behavior," Victor Fuchs, ed., and published as a special issue of *The Journal of Human Resources; Factors in Business Investment*, Robert Eisner; *The Financial Effects of Inflation*, Phillip Cagan and Robert Lipsey; *Foreign Trade Regimes and*

Economic Development: Anatomy and Consequences of Exchange Control Regimes, Jagdish Bhagwati, ed.; *Foreign Trade Regimes and Economic Development: Liberalization Attempts and Consequences*, Anne O. Krueger; "Research in Taxation: A Conference of the National Bureau of Economic Research," Michael Boskin, ed., and published as a special issue of the *Journal of Political Economy*, Part 2, April 1978, *86*.

During 1978 the National Bureau arranged for the University of Chicago Press to be its primary publisher, and ceased publication of two quarterly journals, *Explorations in Economic Research*, and *Annals of Economic and Social Measurement*.

In January 1978 the computer research center under the direction of Edwin Kuh was transferred from the National Bureau to the Massachusetts Institute of Technology. In June 1978 the National Bureau moved to its new headquarters at 1050 Massachusetts Avenue, Cambridge, Massachusetts 02138. This office will house research, the Office of the President, and the financial and administrative operations of the National Bureau. The new address of the New York office is 15-19 West 4th Street, Washington Square, New York, New York 10012. The address of the Palo Alto office remains 204 Junipero Serra Blvd., Palo Alto, California 94305. The Washington office was closed in June 1978.

Further information is available in the *Annual Report*, available on request.

CARL F. CHRIST, *Representative*

Report of the Committee on Economic Education

In 1978 the Committee continued work on two projects. One is revision of the Test of Understanding College Economics (*TUCE*); the revised version should be available for use by instructors and researchers in the 1980–81 academic year. Another is finding ways of expanding professional interest in economic education, broadening the research base, and improving the process of dissemination, with particular attention to the *Journal of Economic Education.*

The Committee brought to a close the developmental phase of a program to provide more effective training in teaching for new Ph.D.s in economics. This is marked by the publication of a 438-page *Resource Manual for Teacher Training Programs in Economics* and the availability of a set of related videotapes that can be used to give graduate students more systematic exposure to what is known about improving their teaching skills. The Committee hopes that these materials will encourage the establishment of regular training programs in graduate economics departments, so as to improve the teaching done by their graduate teaching assistants and by their Ph.D.s after they assume full-time teaching positions.

The scope of the Teacher Training Program, sketched out as part of the Committee's "agenda" prepared in 1972, appeared in the May 1973 *Proceedings* (pp. 303–08). The Sloan Foundation provided substantial funding to support the program's development which began early in 1973. A Planning Committee (Arthur L. Welsh, Chair, W. Lee Hansen, Allen C. Kelley, Darrell R. Lewis, and Phillip Saunders) was established to work under the direction of a broader Advisory Group (G. L. Bach, James M. Buchanan, Rendigs Fels, the late R. A. Gordon, Walter W. Heller, and George J. Stigler). The Planning Committee set to work organizing a pilot workshop for young teachers just completing their Ph.D.s at a number of large institutions; the workshop was held in August 1973 at Indiana University-Bloomington, under the auspices of the Joint Council on Economic Education. About thirty-five advanced graduate students, two or more from each institution represented, plus senior professors from most of the same institutions, participated in the workshop.

The papers and materials prepared for the workshop were tested in pilot teaching programs at the University of Minnesota and the University of Wisconsin-Madison in 1973–74. Four additional schools—Florida State, Indiana, Nebraska, and Purdue—were added in 1974–75. Other institutions, among them, Cornell, Duke, Harvard, Illinois, Lehigh, Missouri, and North Carolina, have since established programs. A preliminary assessment of several of these programs appeared in the May 1976 *Proceedings* (pp. 229–39).

In the meantime, development of the materials continued, with revision of the original chapters, preparation of additional chapters and development of videotapes to complement the written materials. Suggestions from participants and participating institutions, and the experiences of Planning Committee members invited to take part in the programs sponsored by the various institutions, greatly assisted in this work.

The completed *Resource Manual* of sixteen chapters and fifteen videotapes prepared under the editorship of Welsh, Saunders, and Hansen, covers a wide range of topics, including lecturing, discussion leading, preparation of examination questions, organization of a course, instructional objectives, learning theory, use of learning aids. The videotapes are keyed to the *Resource Manual,* many of the chapters include exercises, and instructions are provided on how to use the *Resource Manual.* Descriptions of the training programs from several different institutions are also included; these range from semester-long seminars for academic credit to one-day "crash" programs for teaching assistants in particular courses.

The Teacher Training Program materials now available are viewed as provisional; it is anticipated that a revised set of materials will be prepared at some future date to reflect the experience of additional users. To assist in the

ongoing nature of this project, users of the materials are invited to send their reactions to the authors. There is special need for additional videotapes that can be used to illustrate the nature and possible effectiveness of different teaching techniques.

Copies of the printed *Resource Manual* can be ordered through the Joint Council on Economic Education, 1212 Avenue of the Americas, New York, NY 10036. Descriptions of the videotapes and information about how to order them are included in the *Resource Manual* but can also be obtained through the Joint Council on Economic Education.

A second phase of this project, that of encouraging wide dissemination and use of these materials, is now getting under way. One possibility being explored is the establishment of summer workshops to assist interested institutions and faculty members desiring to establish their own Teacher Training Programs.

W. LEE HANSEN, *Acting Chair*